Tilastolliset menetelmät

Martti Holopainen Pekka Pulkkinen

Sanoma Pro Oy
Helsinki

Tiedustelut
Sanoma Pro Oy, Helsinki
puh. 020 391 000
sähköposti:
asiakaspalvelu@sanomapro.fi

www.sanomapro.fi

Kustantaja: Sanoma Pro Oy
Tämä aineisto on digitaalinen
rinnakkaisversio
tuotteesta ISBN 978-952-63-2000-7
Tilastolliset menetelmät

6 kk lisenssi ISBN 978-952-63-2107-3
48 kk lisenssi ISBN 978-952-63-2108-0

Toimitus: Leena Koskua
Ulkoasu: Khora Oy / Jari Koski
Taitto: Tmi Eija Högman / Katja Hakala
Kuvitus: Krista Partti

© Martti Holopainen, Pekka Pulkkinen
ja Sanoma Pro Oy

Kopiointiehdot
Tämä teos on digitaalinen oppikirja,
joka on suojattu tekijänoikeuslailla
(404/61).

Käyttäjälle

Tämä kirja on kirjoitettu korkeakoulujen perusoppikirjaksi. Erityisen hyvin se sopii ammattikorkeakouluissa opiskeleville ja itseopiskelijoille, jotka tarvitsevat perustiedot tilastollisista menetelmistä. Kirjassa esitellään tilastomenetelmien perusasioita ja painotetaan sellaisia sovelluksia, joista on hyötyä käytännön elämässä. Menetelmien matemaattisen taustan selvittely on jätetty vähäiseksi. Kirjan matemaattinen taso on sen vuoksi pyritty pitämään kohtuullisena, vaikkakin tarpeelliseksi katsottu taustatieto esitetään. Laajan aihepiirinsä vuoksi kirjaa voidaan pitää myös käsikirjana tutkimuksen teosta.

Kirjan esimerkit ja tehtävät ovat pääosin käytännönläheisiä ja tuttuja jokapäiväisestä elämästä. Opiskelija saa tarpeellista harjoitusta käymällä läpi esimerkkejä ja tehtäviä. Koska asioita ei opi pelkästään katselemalla muiden tekemisiä, tehtävien itsenäisen ratkaisemisen tärkeyttä ei voi liikaa korostaa. Tilastolliset periaatteet ja menetelmät oppii ja ymmärtää helpommin, kun ne opettelee itsenäisesti käytännön esimerkkien avulla. Myös ohjelmien käyttö vaatii harjoittelua.

Tilasto-ohjelmat tai vastaavat ovat muuttaneet tilastomenetelmien opiskelua ja oppimista. Ohjelmien avulla on helppo tehdä kokeiluja ja nähdä eri tekijöiden vaikutus. Hyvä ohjelmapaketti on portti tehokkaaseen tilastolliseen analyysiin. Ohjelmien opettaminen ei kuitenkaan kuulu varsinaisesti tämän kirjan aihepiiriin. Ohjelmat opitaan taitavan opettajan johdolla, välillä yrityksen ja erehdyksen kautta. Kirjan esimerkeissä on käytetty useita tietokoneohjelmia: Exceliä, MegaStatia, SPSS:ää, Survoa ja Visual Statisticsia. Kirjasta on pyritty tekemään ohjelmistoriippumaton: kirjassa esitettävät tulostukset on mahdollista toistaa useimmilla tilasto-ohjelmilla tai jopa taulukkolaskentaohjelmilla. Toisaalta kirjaa ei ole tehty ”reseptikirjaksi”, jonka perusteella muutamaa näppäintä painamalla tietokoneesta voisi saada tulosteen, jonka voi liittää tutkimusraporttiin ilman johtopäätösten tekoa ja tulkintaa.

Lukujen alussa olevan ingressitekstin tarkoituksena on johdatella aiheeseen ja luoda kiinnostus kyseisen luvun sisältöön. Tarkoitus on myös motivoida opiskelijoita huomaamaan tilastomenetelmien tärkeys. Jokainen tarvitsee tilastomenetelmien perusasioita esimerkiksi mielipidemittausten ymmärtämiseen, talousraporttien lukemiseen, tutkimustulosten ymmärtämiseen ja kuvanlukutaitoon.

Ensimmäisen luvun tarkoituksena on antaa yleiskuva tieteestä ja tutkimuksenteosta. Kaikilla ei ole riittävää kokonaisnäkemystä ja taustatietoja näistä asioista, joten tämän puolen tuominen esille on tärkeää. Ihminen on usein muiden esittämän informaation varassa ja joutuu tekemään päätöksiä sen perusteella. Joillakin päätöksillä saattaa olla huomattavia taloudellisia vaikutuksia, ja toiset voivat muuten vain olla kauaskantoisia. On siis tärkeätä, että osaamme suhtautua kriittisesti meille esitettyyn informaatioon. Kriittinen suhtautuminen taas on helpompaa, jos osaamme itse hankkia tietoja ja tiedämme, kuinka raakamateriaalista voidaan jalostaa informaatiota, jonka perusteella tulkinnat, arviot ja johtopäätökset tehdään. Toisessa luvussa tutustutaan tilastoaineiston keräämiseen. Samalla käy ilmi, että otoksen poiminta on vaativa tehtävä, joka vaatii aikaa ja osaamista ja aiheuttaa kustannuksia. Niille, jotka keräävät tilastoaineiston kyselylomakkeilla, annetaan luvun lopussa lyhyesti ohjeita lomakkeen laatimista varten.

Kirjan kolmannessa ja neljännessä luvussa perehdytään tilastojen kuvaamiseen grafiikkaa ja tunnuslukuja käyttämällä. Tilastojen parissa työskentelevän pitäisi pystyä muuntamaan data käyttökelpoiseksi informaatioksi ja tulkitsemaan se, ja usein asioita voidaan havainnollistaa graafisin kuvin. Hyvän, oikein tehdyn kuvan tekeminen on oikeastaan oma taiteenlajinsa. Tässä kirjassa tyydytään kuitenkin vain esittelemään perustapaukset.

Tunnusluvut saadaan helposti tilasto-ohjelmien avulla taulukkona, josta on helppo valita sopivat tunnusluvut. Joitakin esimerkkejä on laskettu laskimella, mutta muutoin laskut suoritetaan tietokoneohjelmalla. Kaavoilla laskeminen on otettu kirjassa mukaan silloin, kun se auttaa ymmärtämään perusideaa. Ongelmana ei olekaan enää itse tunnuslukujen laskeminen, vaan keskeistä on niiden merkityksen ymmärtäminen ja tulkinta.

Todennäköisyyden käsitettä käytetään jatkuvasti esimerkiksi sääennusteissa ja pörssikurssien analyyseissä. Todennäköisyyslaskenta onkin tilastomenetelmien kivijalka, ja ilman sen periaatteiden tuntemista johtopäätösten luotettavuustarkastelut jäävät ymmärtämättä. Kuudennen luvun alussa puhutaan todennäköisyydestä ja luvun loppuosassa tarkastelua laajennetaan todennäköisyysjakaumiin.

Päätöksenteon pitää perustua faktoihin, ei luuloihin, toiveisiin tai pelkoihin. Päätöksenteon perustana on kulloinkin käytettävissä oleva tieto. Päätöksentekoa helpottamaan on kehitetty useita menetelmiä, joista monet ovat tilastollisia. Päätöksenteossa voidaan myös joutua harhaan, jos esimerkiksi käytetään vääriä menetelmiä tai tehdään vääriä johtopäätöksiä. Myös väärän johtopäätöksen tekemisen riski on otettava huomioon. Riskin suuruus

ilmoitetaan yleensä todennäköisyytenä. Monessa tapauksessa päätöksenteossa käytetyn riskin suuruuteen voidaan itse vaikuttaa pelkästään valitsemalla alhaisempi riskitaso. Kun riskin todennäköisyys on tiedossa, kyseessä on hallittu riski. Loogisten johtopäätösten hallittu tekeminen vaatiikin aina tilastollisten perusmenetelmien, niiden oikean käytön ja käytön riskien tuntemista.

Koko tilastotieteen kannalta tärkeään ja käytännönläheiseen parametrien estimointiin ja hypoteesien testaukseen perehdytään kirjan seitsemännessä luvussa. Tämä laaja luku valottaa testausta käytännön kannalta. Nykyisin on onneksi käytettävissä hyviä tilasto-ohjelmia, joilla mekaaniset laskutoimitukset on vaivatonta tehdä. Tekninen laskenta suoritetaan pääasiassa ohjelmien avulla, mutta tulkinta ja johtopäätökset jäävät ohjelman käyttäjän tehtäväksi.

Kahdeksannessa luvussa tutkitaan tilastollista riippuvuutta sekä graafisesti että laskennallisesti. Riippuvuutta voidaan mitata korrelaatiokertoimien avulla. Lisäksi luvun loppupuolella selvitellään niiden merkitsevyyttä ja tulkintaa.

Yhdeksännessä luvussa tarkastellaan regressiota. Aluksi pyritään löytämään matemaattinen malli, jonka avulla voidaan laatia ennusteita. Luvun loppuosassa puolestaan tutkitaan mallin hyvyyttä ja ennusteiden luotettavuutta.

Kirjan kymmenes luku käsittelee aikasarjoja, niiden komponentteja, kuvaamista ja mallintamista. Esitetyt menetelmät soveltuvat kuitenkin vain melko säännöllisille aikasarjoille. Epäsäännöllisempien aikasarjojen käsittely vaatii tarkempia menetelmiä, jotka eivät kuulu tämän esityksen piiriin.

Kirjassa käytetyt aineistot ovat saatavilla internetissä kustantajan verkko-osoitteessa www.sanomapro.fi > Oppilaitoksille > Ammatti- ja korkeakoulukirjat >Ammattiverkko.

Kirjoitustyö jakaantui tekijöiden kesken siten, että Martti Holopainen on kirjoittanut luvut 1, 2 ja 8–10 sekä esimerkit B, C ja D ja Pekka Pulkkinen luvut 3–7 sekä esimerkin A.

Muutoksia 5. uudistettuun painokseen

Kirjaan on tehty koko joukko uudistuksia. Uudistusten virikkeinä ovat olleet kirjasta saatu palaute ja oman opetuksen tuoma kokemus. Lisäksi on otettu huomioon ja käyttöön ASA:n (American Statistical Association) suositukset tilastomatematiikan opetuksesta. Niinpä esimerkiksi tietokoneen käyttöä apuvälineenä on entisestään lisätty. Samoin tilastollisten termien englanninkielisiä vastineita on lisätty, sillä monet tilasto-ohjelmat ovat englanninkielisiä. Opiskelijat joutuvat muissakin opinnoissaan käyttämään lähdekirjallisuutena englanninkielisiä kirjoja, joten englanninkielisten termien tunteminen on välttämätöntä.

Kirjan tekstiä on muokattu ja tarkennettu. Muutoksissa on otettu huomioon ammattikorkeakouluopiskelijoiden tarpeet ja matematiikan taitojen todellinen taso, mikä on vaatinut enemmän selittävää tekstiä. Tekstin muokkaus on vaatinut myös kaikkien aikaan sidottujen ja aikaan liittyvien tilastojen ajantasaistamista, ja kirjaan on lisätty uusia esimerkkejä ja tehtäviä muiltakin kuin liiketalouden alalta.

Edellisissä painoksissa kirjan lopussa olevat esimerkkitapaukset on nyt sijoitettu asianomaisten lukujen yhteyteen, jolloin ne antavat laajemman kuvan luvussa käsiteltyjen menetelmien käytöstä.

Kirjaan on saatavana laaja opettajanaineisto Sanoma Pron verkkokaupasta.

Valkeakoskella ja Kuopiossa tammikuussa 2008

Martti Holopainen *Pekka Pulkkinen*

SISÄLTÖ

1 Tiede, tutkimus ja tilastollinen ajatteluprosessi 13
1.1 Tiete ja tieteellinen menetelmä. 13
1.2 Mittaaminen, validiteetti ja reliabiliteetti. 15
1.3 Tutkimusprosessi 18
1.4 Tutkimustyyppejä 20
1.5 Tilastot ja internetin käyttö 22
Tilastot 22
Internet 23
Tiedon omakohtainen kerääminen 25
1.6 Tilastollinen ajatteluprosessi 25
Tehtäviä 27

2 Tilastoaineiston kerääminen 28
2.1 Kokonaistutkimus ja otantatutkimus 29
Otantatutkimuksen vaiheet 30
Otantatutkimus vai kokonaistutkimus? 30
2.2 Otantamenetelmiä 31
Yksinkertainen satunnaisotanta 31
Systemaattinen otanta 32
Ositettu otanta 33
Ryväsotanta 35
Harkinnanvarainen otanta 36
Kiintiöotanta 36
2.3 Otoskoon merkitys 37
Virhemarginaalit 37
Otoskoko suhteellista osuutta määritettäessä 40
2.4 Otantatutkimuksen virhelähteitä 41
2.5 Kyselylomakkeet 42
Tehtäviä 44

3 Tilastojen tutkiminen 46
3.1 Luokittelu ja taulukointi 46
Havaintomatriisi 46
Ristiintaulukointi 52
3.2 Tilastojen graafinen esittäminen 53
Pylväsdiagrammit 53
Viivadiagrammit 61
Sektoridiagrammit 63
Muita kuvioita 65
Grafiikka voi myös vääristää totuutta 67
Tehtäviä 70

4 Suoran jakauman tunnusluvut 78
4.1 Sijaintiluvut 79
Keskiluvut 79
Moodi 79
Mediaani 80
Aritmeettinen keskiarvo 83
Muita sijaintilukuja 86
4.2 Hajontaluvut 88
Vaihteluväli ja kvartiiliväli 89
Vaihteluvälin pituus ja kvarttiilipoikkeama 90
Keskihajonta 90
Variaatiokerroin 93
4.3 Muita tunnuslukuja 94
Vinous ja huipukkuus 94
Tehtäviä 98

5 Todennäköisyyslaskenta 102
5.1 Kombinaatio-oppia 102
Tuloperiaate 104
Permutaatio 104
Kombinaatio 106
5.2 Todennäköisyyslaskennan peruskäsitteitä 107
Subjektiivinen lähestymistapa 107
Klassinen lähestymistapa 108
Empiirinen lähestymistapa 109
Tilastollinen lähestymistapa 111
Aksiomaattinen lähestymistapa 113

5.3 Todennäköisyyden laskusääntöjä 113
Erillisten tapahtumien yhteenlaskusääntö 114
Kahden tapahtuman yleinen yhteenlaskusääntö 115
Riippumattomien tapahtumien kertolaskusääntö 116
Yleinen kertolaskusääntö 117
Komplementti 118
Tehtäviä 119

6 Todennäköisyysjakaumat 124
6.1 Diskreetit todennäköisyysjakaumat 125
Satunnaismuuttuja 125
Diskreetti todennäköisyysjakauma 126
Diskreetin jakauman kertymäfunktio 128
Diskreetin jakauman kertymäfunktion ominaisuuksia 130
Diskreetin jakauman odotusarvo 131
Keskihajonta 132
Binomijakauma 132
Binomijakauman todennäköisyysfunktio 133
Binomijakauman kertymäfunktio 135
Binomijakauman odotusarvo ja keskihajonta 137
Poisson-jakauma 137
Binomi- ja Poisson-jakaumien yhteys 138
Poisson-jakauman pistetodennäköisyysfunktio 139
6.2 Jatkuvat todennäköisyysjakaumat 141
Tiheysfunktion ominaisuuksia 142
Jatkuvan jakauman kertymäfunktio 142
Jatkuvan jakauman kertymäfunktion ominaisuuksia 143
Normaalijakauma 144
Normaalijakauman määritelmä 145
Normaalijakauman kertymäfunktio 146
Normitettu satunnaismuuttuja 146
Normaalijakauman ominaisuuksia 148
Keskeinen raja-arvolause 152
Binomi- ja normaalijakauman yhteys 153
Keskiarvon otantajakauma 154
Studentin t-jakauma 155
χ^2-jakauma 156
Tehtäviä 157

7 Tilastollinen päätöksenteko 164
Tilastollinen päättely 165
7.1 Parametrien estimointi 165
Estimointi 165
Luottamusväli 166
Odotusarvon luottamusväli 168
Prosenttiosuuden luottamusväli 170
Otoskoko 173
Virhemarginaali 174
7.2 Hypoteesien testaus 175
Tilastollisen päättelyn vaiheet 176
7.3 Parametriset testit 178
Keskiarvotesti 181
Studentin t-testi 182
Kahden riippumattoman otoksen keskiarvotesti 185
Kahden riippumattoman otoksen keskiarvojen vertailu t-testin avulla 187
Verrannollisten parien t-testi 190
Suhteellisen osuuden testaus 192
Suhteellisen osuuden testaus kahdesta otoksesta 193
7.4 Jakaumasta riippumattomat testit 195
Mann-Whitneyn testi 197
Wilcoxonin testi 198
χ^2-yhteensopivuustesti 200
χ^2-riippumattomuustesti 202
7.5 Testin valinta 204
7.6 Case A: Tavaratalon palvelutasotutkimus 207
Tehtäviä 219

8 Tilastollinen riippuvuus 228
8.1 Hajontakuviot 228
Perustapaus 228
Samanaikaiset hajontakuviot 230
8.2 Korrelaatio 233
Pearsonin korrelaatiokerroin 233
Korrelaatiokertoimen ominaisuuksia 234
Korrelaatiokertoimen laskeminen 235
Korrelaatiomatriisi 239
Spearmanin järjestyskorrelaatiokerroin 239
8.3 Korrelaatiokertoimen merkitsevyys 242
8.4 Kertoimien tulkinta 245
8.5 Yhteyden tutkimiseen liittyvä vuokaavio 248
8.6 Case B: Pakokaasujen pitoisuudet 249
Tehtäviä 252

9 Regressio 259
9.1 Pienimmän neliösumman menetelmä 259
9.2 Lineaarinen malli 261
Regressiosuoran yhtälö 261
Tietokonetulosteen tulkinta 265
Ennusteiden laatiminen 266
9.3 Muita malleja 268
Eksponentiaalinen malli 269
Polynomimalli 273
9.4 Usean muuttujan mallit 275
9.5 Mallin hyvyys 277
Selityskerroin 277
T-arvot 278
Jäännöstermin merkitys 282
Diagnostiikka 283
9.6 Ennusteen luotettavuus 288
Graafinen tarkastelu 288
Ennusteväli ja ennustevyöhyke 289
9.7 Regressiomallin vuokaavio 291
9.8 Case C: Polttoaineen kulutus 293
Tehtäviä 301

10 Aikasarjat 305
10.1 Kuvaajat ja komponentit 305
Aikasarjan kuvaajat 306
Aikasarjan komponentit 311
Komponenttien yhdistäminen 314
10.2 Trendi 316
Trendin määrittäminen 316
Lineaarisen trendin olemassaolon testaus 318
10.3 Kausivaihtelu 319
Kausivaihtelun olemassaolo 319
Liukuvan keskiarvon käyttö 320
Kausivaihteluindeksien laskeminen 322
10.4 Suhdannevaihtelu 325
10.5 Satunnaisvaihtelu ja jäännössarja 327
10.6 Ennusteiden laatiminen 327
Ennusteen laatiminen tietokoneohjelmalla 330
Ennusteiden luotettavuus 330
10.7 Case D: Vanhuuseläkeläisten määrän kehittyminen 332
Tehtäviä 339

Liitteet 344
Tehtävien vastauksia 344
Normitettu normaalijakauma 350
T-jakaumaan liittyviä kriittisiä arvoja 352
X^2-jakauman kriittisiä arvoja 353
Pearsonin korrelaatiokertoimen kriittisiä arvoja 354
Spearmanin järjestyskorrelaatiokertoimen kriittisiä arvoja 355
Hakemisto 356

1
Tiede, tutkimus ja tilastollinen ajatteluprosessi

Ihmisellä on luontainen pyrkimys ymmärtää, mitä hänen ympärillään tapahtuu. Ymmärrys perustuu ihmisen tekemiin havaintoihin, joita luokittelemalla tai seuraamalla hän pyrkii löytämään säännönmukaisuuksia. Näiden säännönmukaisuuksien löytäminen vaatii loogisten johtopäätösten tekoa. Pelkän uteliaisuuden tyydyttämisen ja älyllisen mielihyvän lisäksi ihminen pyrkii ennakoimaan tulevaa ja siten varustautumaan tuleviin tapahtumiin.

Edellä kuvattuja taitoja voi oppia. Tämän kirjan tarkoitus onkin näyttää, kuinka havainnoista edetään havaintotulosten ymmärtämiseen ja johtopäätösten tekemiseen.

1.1 Tiede ja tieteellinen menetelmä

Tiede (science) muodostuu kahdesta komponentista: tieteellisestä tutkimusprosessista ja tieteellisen tutkimuksen tuottamista tuloksista. Tieteen tulokset saavutetaan **tieteellisten menetelmien** (scientific methods) avulla.

Tieteelliseltä menetelmältä vaaditaan

- loogisuutta
- objektiivisuutta
- yleistettävyyttä
- todennettavuutta
- yhteensopivuutta teorian ja havaintojen kanssa.

Tieteellisten menetelmien avulla pyritään systemaattisesti keräämään ihmistä, luontoa ja yhteiskuntaa koskevaa tietoa. Pelkän tiedon kerääminen ei kuitenkaan riitä vaan tiedoista pitää osata tehdä oikeat **johto-**

päätökset (inferences). Aikojen kuluessa havaintojen ja johtopäätösten tekoa on pyritty systematisoimaan niin, että loogisten johtopäätösten tekeminen helpottuisi. Osa tieteellisistä menetelmistä keskittyykin juuri johtopäätösten tekemiseen.

Johtopäätökset tehdään tavallisesti joko **deduktiivisesti** tai **induktiivisesti**. Deduktiivisessa päättelyssä edetään yleisestä tilanteesta yksittäiseen tapaukseen, siis vaikkapa kuluttajakäyttäytymistä koskevan teorian soveltamisesta yksittäiseen kuluttajaan. Tällöin yleisen teorian pohjalta laaditaan ennusteita yksittäisten kuluttajien (keskimääräiselle) käyttäytymiselle.

Induktiivisessa päättelyssä taas edetään yksittäistapauksesta yleistykseen. Induktiivinen päättely perustuu siihen olettamukseen, että uudet havainnot ovat yleistysmielessä samantapaisia kuin vanhat havainnot tai että tulevaisuus noudattelee menneisyyttä. Vaarana kuitenkin on, että tehdään vääriä johtopäätöksiä. Uudet havainnot saattavat tukea aikaisempia käsityksiä, mutta vaikka uusia havaintoja kerätään paljonkin, niillä on vaikea sitovasti todistaa yleistystä. Vanha esimerkki osoittaa tämän todeksi: keväisin ja syksyisin näkemämme valkoiset joutsenet antavat aiheen olettaa, että kaikki joutsenet ovat valkoisia. Jokainen uusi näkemämme valkoinen joutsen vahvistaa käsitystämme siitä, että kaikki joutsenet ovat valkoisia, mutta silti tämä ei todista väitettämme todeksi. Yksikin havaittu musta joutsen pystyy osoittamaan yleistyksemme vääräksi.

Tieteen edistämiseksi tarvitaan sekä deduktiivista että induktiivista päättelyä. Uudet **otaksumat** (conjectures) tai **mallit** syntyvät usein induktiivisen päättelyn tuloksena. Jos otaksuma tai malli osoittautuu oikeaksi, sen perusteella voidaan deduktion avulla ennakoida tulevaisuutta. Jos taas otaksuma tai malli osoittautuu vääräksi, sitä on muokattava tai se on korvattava kokonaan uudella.

Tieteessä pyrkimyksenä on koota ja systematisoida aikaisempia tutkimustuloksia. Tämän työn tuloksena voi syntyä uusia kattavia teorioita, kuten suhteellisuusteoria. Pienemmässä mittakaavassa tuloksena saattaa olla otaksuma tai jotakin ilmiötä kuvaava malli. Teoria, otaksuma, malli tai tutkimustulos on kuitenkin testattava ennen kuin se voidaan hyväksyä. Usein testaus tehdään tilastollisesti. Tilastolliseen testaamiseen palataan lähemmin kirjan 7. luvussa.

1.2 Mittaaminen, validiteetti ja reliabiliteetti

Mitä tahansa tutkimusta tehtäessä aluksi on määriteltävä, mikä joukko on tutkimuksen kohteena. Tätä joukkoa sanotaan **perusjoukoksi** (population), ja se muodostuu **tutkimusyksiköistä,** joita tilastollisissa tutkimuksissa sanotaan **tilastoyksiköiksi**.

Tilastoyksiköistä kerätään tietoja mittaamalla. **Mittaamisella** tarkoitetaan tapahtumaa, jossa tilastoyksikön tarkasteltavaan ominaisuuteen liitetään mittaluku tai mittasymboli. Mittaamisen tasoa kuvataan mitta-asteikoilla, joita ovat

- luokitteluasteikko
- järjestysasteikko
- välimatka-asteikko
- suhdeasteikko.

Luokitteluasteikolla (nominal scale) tilastoyksiköt luokitellaan vain ennalta määrättyihin luokkiin, jolloin luokkien järjestyksellä ei ole merkitystä. Luokittelu tulee tehdä siten, että kukin tilastoyksikkö kuuluu vain yhteen luokkaan. Tällöin kahdesta tilastoyksiköstä voidaan päätellä vain, kuuluvatko ne samaan luokkaan vai eivät. Esimerkkejä luokitteluasteikon tasoisesta mittauksesta ovat sukupuolen, veriryhmän, uskonnollisen vakaumuksen ja kotikunnan mittaus. Vaikka luokkia voidaan kuvata myös numeroarvoilla, niillä ei voi tehdä mitään laskutoimituksia.

Järjestysasteikolla (ordinal scale) tilastoyksiköt luokitellaan ennalta määrättyihin luokkiin, joilla on yksikäsitteinen järjestys. Esimerkkejä tästä ovat sosiaaliryhmä, sotilasarvot, kauneuskilpailujen tulos ja monet mielipidemittaukset. Vaikka luokkia nytkin voidaan kuvata numeroarvoilla, niillä ei yleisesti ottaen pidä tehdä mitään laskutoimituksia. Monissa mielipidemittauksissa on kuitenkin tapana laskea muun muassa keskiarvoja. Tällöin kyse on enemmänkin yleiskuvan saamisesta kuin aritmeettisen laskutoimituksen antamasta tarkasta tuloksesta.

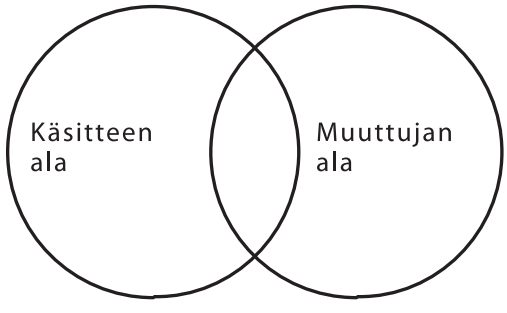
Välimatka-asteikolla (interval scale) luokat ovat samanlevyisiä ja välimatka luokasta toiseen on yhtä suuri asteikon joka kohdassa. Välimatka-asteikolla pystytään mittaamaan yksittäisten luokkien tai havaintoarvojen ero. Tyypillinen tilanne on lämpötilan ilmoittaminen celsius- tai fahrenheitasteina. Pystymme numeroarvoina ilmoittamaan, onko tänään lämpimämpi, yhtä lämmin vai kylmempi sää kuin eilen ja kuinka monta astetta muutos on. Havaintoarvoilla on mielekästä suorittaa laskutoimituksia.

Suhdeasteikolla (ratio scale) on absoluuttinen nollapiste, ja sillä on mielekästä laskea myös lukujen suhteita. Esimerkiksi kuuden euron hintainen tuote on kaksi kertaa niin kallis kuin kolmen euron hintainen tuote. Esimerkkejä suhdeasteikon tasoisista mittauksista ovat perheen lasten lukumäärän, kaupungin asukasluvun, välimatkojen, yrityksen liikevaihdon sekä tuotteen hinnan ja painon mittaukset. Suhdeasteikolla kaikki laskutoimitukset ja asteikkomuunnokset ovat mahdollisia.

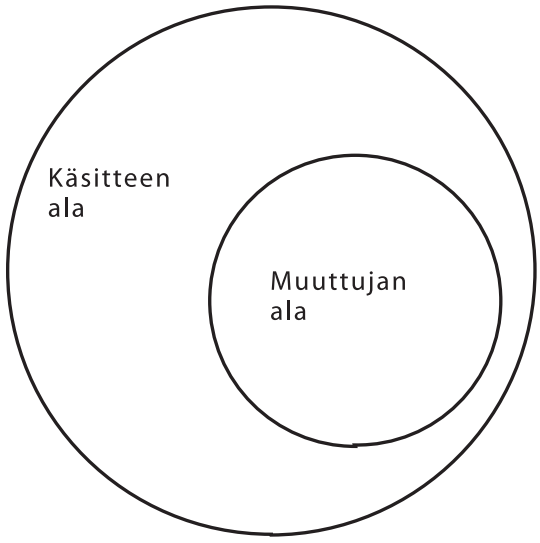
Mitta-asteikkojen tunnistamisella on käytännön merkitystä muun muassa perusjoukkoa kuvaavien tunnuslukujen määrittämisessä, sillä mitta-asteikko voi rajoittaa tunnuslukujen käyttöä. Samoin havaintoaineistoa graafisesti esitettäessä täytyy mitta-asteikko ottaa huomioon.

Kun tilastoyksiköistä kerätään tietoja mittaamalla, mittaaminen kohdistuu johonkin **muuttujaan**. Muuttujat voidaan luokitella esimerkiksi **kvalitatiivisiin** ja **kvantitatiivisiin** muuttujiin. Kvalitatiivisiin muuttujiin liittyy luokittelu- tai järjestysasteikko ja kvantitatiivisiin muuttujiin välimatka- ja suhdeasteikko. Muuttujat voidaan luokitella myös **diskreetteihin** ja **jatkuviin** muuttujiin. Diskreetti muuttuja voi saada vain yksittäisiä arvoja, kun taas jatkuva muuttuja voi saada tietyllä välillä kaikki mahdolliset arvot.

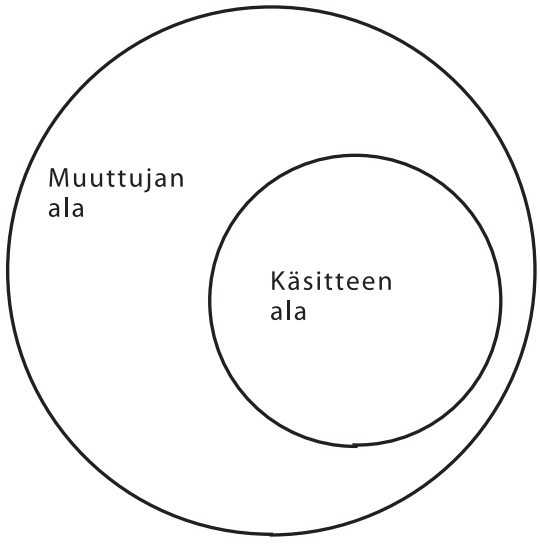
Mittaamisen **validiteetti** ilmaisee, missä määrin on kyetty mittaamaan juuri sitä, mitä pitikin mitata. Ideaalitilanne on se, että muuttuja mittaa täsmälleen sitä, mitä oli tarkoitus mitata. Näin ei kuitenkaan aina ole. Tällöin voi esiintyä seuraavia validiteettia heikentäviä systemaattisia virhetilanteita:



Muuttujan ala ja käsitteen ala kohtaavat vain osittain.



Muuttujan ala on rajoittuneempi kuin käsitteen ala.



Muuttujan ala on laajempi kuin käsitteen ala.

Mittauksen **reliabiliteetilla** tarkoitetaan mittarin luotettavuutta eli kykyä tuottaa ei-sattumanvaraisia tuloksia. Mittauksen reliabiliteetti on suuri, jos eri mittauskerroilla saadaan samanlaisia tuloksia samasta tai samantapaisesta aineistosta.

1.3 Tutkimusprosessi

Tutkimus (research, study) on systemaattista toimintaa tiedon lisäämiseksi. Empiirinen tutkimusprosessi voidaan karkeasti jakaa

- tutkimuksen suunnitteluun
- tutkimusongelman asettamiseen ja tavoitteiden määrittelyyn
- tutkimusstrategian laatimiseen
- aineiston keräämiseen
- aineiston kuvaamiseen
- aineiston analyysiin
- tulkintaan ja johtopäätösten tekoon
- arviointiin
- raportin laatimiseen.

Tutkimuksen suunnittelun yhteydessä selvitetään, **mitä aiotaan tehdä** eli millainen kysymys, ongelma tai seikka halutaan selvittää. Tutkimuksen suunnitteluun saattaa liittyä esitutkimuksen tekemistä. **Esitutkimuksessa** tutustutaan aikaisempiin vastaavanlaisiin tutkimuksiin, tutkitaan aiheeseen liittyvää kirjallisuutta ja pohditaan tutkimuksen tekotapaa.

Tutkimusongelman asettamisessa ja tavoitteiden määrittelyssä tarkennetaan ja selvästi yksilöidään, **mitä tullaan tekemään**. Ongelman yksityiskohtainen hahmottaminen ja tavoitteiden selkeä asettaminen on tutkimuksen onnistumiseksi keskeistä. Jos tämä vaihe laiminlyödään tai tehdään huolimattomasti, tietojen hankinta on sattumanvaraista. Tällöin tutkimuksen lopputulos riippuu paljolti siitä, mitä tietoja on satuttu saamaan tai keräämään.

Tutkimusstrategian laatimisen yhteydessä päätetään, millainen tutkimus tehdään: onko kyseessä esimerkiksi kuvaileva tutkimus asiaintilan selvittämiseksi vai kausaalinen tutkimus syy–seuraussuhteiden selvittämiseksi. Tutkimusstrategian laatimisen yhteydessä päätetään myös, **miten tutkimus lopulta tehdään** eli miten tavoitteisiin päästään.

Tällöin lyödään lukkoon muun muassa
– aikataulu
– rahoitus
– tutkimuksen suorittajat ja vastuu
– suoritetaanko kokonaistutkimus vai otantatutkimus
– otantamenetelmä
– otoskoko
– aineiston keräystapa
– (tilastollinen) menetelmä.

Kun edellä mainitut asiat on lyöty lukkoon, samalla on jo päätetty, mitä ja miten tarkkoja tuloksia tutkimuksella on mahdollista saada. Tämän vuoksi tutkimusstrategian laadinnan on oltava tarkkaa, ettei myöhemmin tehtävä työ menisi hukkaan.

Aineistoa voidaan kerätä esimerkiksi kyselylomakkeilla, mittaamalla, havainnoimalla tai käyttämällä valmiita tilastoja. Aineiston keräämiseen liittyviä asioita käsitellään luvussa 2 otantamenetelmien yhteydessä.

Aineiston kuvaaminen käsittää aineiston tarkastelun, luokittelun ja graafisen esittämisen. Sama aineisto voidaan luokitella, kuvata taulukkona ja esittää graafisesti monellakin eri tavalla. Eri vaihtoehtoja tutkimalla ja vertaamalla löydetään kuhunkin tilanteeseen parhaiten sopiva kuvaamistapa. Näihin asioihin tutustutaan luvussa 3.

Aineistoa voidaan analysoida monella tavalla. Analyysin valintaan vaikuttaa ratkaisevasti se, millainen tutkimus on kyseessä. Jos kyseessä on tilastollinen tutkimus, tilastomenetelmiä on sekä kvalitatiiviseen että kvantitatiiviseen tutkimukseen. Tässä kirjassa tutustutaan tavallisimpiin tilastollisiin menetelmiin.

Tulkinta (interpretation) on keskeisellä sijalla kaikessa tutkimustoiminnassa. Tulkinnan yhteydessä tutkimuksen löydökset muuttuvat merkityksiksi. Tämän vuoksi tulkinta ja johtopäätösten teko ovat keskeisellä sijalla tässäkin kirjassa. Jos analyysin tuloksia ei osata tulkita eikä niiden merkitystä nähdä, pääosa tutkimuksesta on mennyt hukkaan.

Ennen kuin tutkimus on lopullisesti tehty, on hetkeksi pysähdyttävä arvioimaan käytettyjä menetelmiä ja tuloksia. Tällöin on muun muassa pystyttävä katsomaan kriittisesti aikaisempia tutkimustuloksia ja omia tuloksia. Jos syntyy selkeitä eroja aikaisempiin tutkimuksiin, on pyrit-

tävä arvioimaan, mistä erot voisivat johtua: Onko käytetty tilanteeseen sopivia menetelmiä? Onko aineistoa riittävästi luotettavien johtopäätösten tekemiseen? Aiheuttaako mahdollinen kato tulosten vääristymistä? Koskevatko tehdyt johtopäätökset koko perusjoukkoa?

Vielä kerran on muistettava myös lähdekritiikki eli pohdittava, ovatko kaikki käytetyt lähteet kaiken nykytietämyksen valossa luotettavia ja riittävän monipuolisia. Lopuksi tutkimuksesta kirjoitetaan raportti, jossa esitellään tutkimusprosessin vaiheet, perustellaan menetelmien käyttö, selostetaan löydökset ja niiden merkitykset selväkielisesti, tehdään johtopäätökset sekä arvioidaan tuloksia, niiden luotettavuutta ja merkitystä.

1.4 Tutkimustyyppejä

Vaikka tässä kirjassa käsitellään lähinnä tilastollisia tutkimuksia, olisi väärin luulla, että kaikki tutkimukset olisivat tilastollisia tutkimuksia. Itse asiassa kaikista tutkimuksista suurin osa lienee muita kuin tilastollisia tutkimuksia, joskin tilastollisia osuuksia saattaa sisältyä näihinkin tutkimuksiin.

Tutkimukset voidaan jakaa **teoreettisiin** ja **empiirisiin** tutkimuksiin. Teoreettinen tutkimus on pitkälti **perustutkimusta,** jonka tarkoituksena on uuden tieteellisen tiedon etsintä ja uusien menetelmien luonti ilman, että nimenomaisena päämääränä ovat käytännön sovellukset. Empiiriset tutkimukset taas ovat **soveltavia tutkimuksia**, ja niillä tarkoitetaan käytännön tavoitteisiin pyrkivää itsenäistä ja omaperäistä tiedon etsintää, jossa nojaudutaan perustutkimuksen tuloksiin.

Empiiriset tutkimukset voidaan jakaa esimerkiksi seuraaviin kahteen ryhmään:

- kvalitatiivisiin tutkimuksiin
- kvantitatiivisiin tutkimuksiin.

Kvalitatiiviset tutkimukset liittyvät läheisesti esimerkiksi sosiaalitieteisiin, käyttäytymistieteeseen, kasvatustieteeseen, ekologiaan, markkinointiin ja terveydenhuoltoon, joissa tutkittavia ominaisuuksia mitataan usein kvalitatiivisilla (laadullisilla) muuttujilla. Tutkimuksilla

pyritään saamaan vastaus kysymyksiin ”miksi?”, ”miten?” ja ”millainen?”, ja tutkimusaineistot ovat usein varsin suppeita.

Kvantitatiiviset tutkimukset taas tulevat kysymykseen silloin, kun tutkittavan ominaisuuden mittaamisessa käytetään välimatka- tai suhdeasteikkoa, ja niillä pyritään saamaan vastaus kysymyksiin ”mikä?”, ”missä?”, ”kuinka usein?” ja ”kuinka paljon?”. Tutkimusaineistot ovat usein suuria, ja ilmiöitä kuvataan numeerisesti. Usein luullaan virheellisesti, että tilastollisia menetelmiä voidaan käyttää vain kvantitatiivisissa tutkimuksissa, mutta tilastollisia menetelmiä on kehitetty myös kvalitatiivista tutkimusta varten.

Empiiriset tutkimukset voidaan jakaa myös **havainnoiviin** (observational) ja **kokeellisiin** (experimental) tutkimuksiin. Havainnoivassa tutkimuksessa ideaalitapaus on sellainen, jossa ilmiön kehittymistä tulevaisuudessa pystytään seuraamaan pitkän aikavälin kuluessa. Tällaista tapausta sanotaan ilmiön **prospektiiviseksi** tutkimukseksi. Jos ilmiön kehittymistä seurataan takautuvasti, tutkimusta sanotaan **retrospektiiviseksi**. Sekä prospektiivinen että retrospektiivinen tutkimus ovat **pitkittäistutkimuksia** (longitudinal study).

Empiirisessä **poikkileikkaustutkimuksessa** (cross-sectional study) tutkitaan tiettynä ajankohtana tietyn ilmiön tai useiden ilmiöiden eri puolia, esimerkiksi kotitalouksien ruokaan kuluttamaa rahamäärää.

Kausaalisessa tutkimuksessa pyritään selvittämään muuttujien välisiä syy–seuraussuhteita.

Kuvailevassa tutkimuksessa esitetään asiaintila joko tiettynä hetkenä tai pitemmän ajan kuluessa.

Kirjoituspöytätutkimukseksi sanotaan tutkimusta, jossa kaikki tieto on valmiiksi kerätty. **Case-tutkimukse**n tarkoituksena on tutkia tietty yksittäistapaus mahdollisimman tarkasti, tehdä sitä koskevat johtopäätökset ja mahdollisesti esittää toimenpidesuositukset tai yleistykset. **Survey-tutkimus** taas on suunnitelmallinen kysely- tai haastattelututkimus, joka perustuu tutkimusaineiston keräämiseen valmiiksi jäsennellyillä lomakkeilla. Kerätty tieto voidaan analysoida useilla eri menetelmillä.

Empiiristä tutkimusta suunniteltaessa on mietittävä, millaista tutkimusstrategiaa kannattaa käyttää. Oheinen taulukko antaa asiasta joitakin viitteitä.

Tutkimusstrategian valintaan liittyviä vaihtoehtoja

Tutkimus-strategia	Tutkimus-kysymyksiä	Pitääkö tapahtumia pystyä kontrolloi-maan?	Kohdistuuko tutkimus nyky-tapahtumiin?
Aikaisemmin tapahtuneen ilmiön tutkiminen	Miten, miksi, milloin, kuka, mikä?	Ei	Ei
Kirjoituspöytä-tutkimus	Kuka, mikä, mitä, missä, milloin, kuinka monta, kuinka paljon?	Ei	Kyllä/Ei
Koejärjestely	Miten, miksi, millainen?	Kyllä	Kyllä
Case-tutkimus	Miten, miksi, millainen?	Ei	Kyllä
Survey-tutkimus	Kuka, mikä, mitä, missä, milloin, kuinka monta, kuinka paljon?	Ei	Kyllä

1.5 Tilastot ja internetin käyttö

Tutkimuksella pyritään ratkaisemaan jokin ongelma tai selvittämään asioiden tila. Tutkimus pohjautuu faktoihin, joita kerätään monella tavalla. Valmista havaintoaineistoa kannattaa tietysti hyödyntää, jos sitä on käytettävissä.

Valmiita tilastoja on runsaasti muun muassa yrityksillä, yhteisöillä, kunnilla ja valtiolla. Usein vaikeutena on pikemminkin oikean tiedon löytäminen kuin se, ettei tietoa olisi. Tutkimuksen pohjatietoja voi saada aikaisemmista tutkimuksista tai alan muista julkaisuista.

Tilastot

Suomen tilastollinen keskusvirasto, **Tilastokeskus**, tuottaa suuren joukon julkisia tilastoja. Tilastokeskuksen kattavin julkaisu on **Suomen tilastollinen vuosikirja**, joka sisältää yhteiskuntaelämän ja valtion eri aloilta hyvin monipuolisen valikoiman sekä koti- että ulkomaisia tilastotietoja. Muita Tilastokeskuksen julkaisuja ovat muun muassa Suomen virallinen tilasto ja Tilastotiedotussarja.

Tilastoja laativat muun muassa valtion ministeriöt ja laitokset sekä tutkimuslaitokset. Lisäksi niitä laativat esimerkiksi Eläketurvakeskus, Väestörekisterikeskus ja Valtion taloudellinen tutkimuskeskus. Myös suurimmilla kaupungeilla on omat tilastolliset vuosikirjansa.

Valmiissa tilastoissa käytetään sovittuja tilastosymboleita

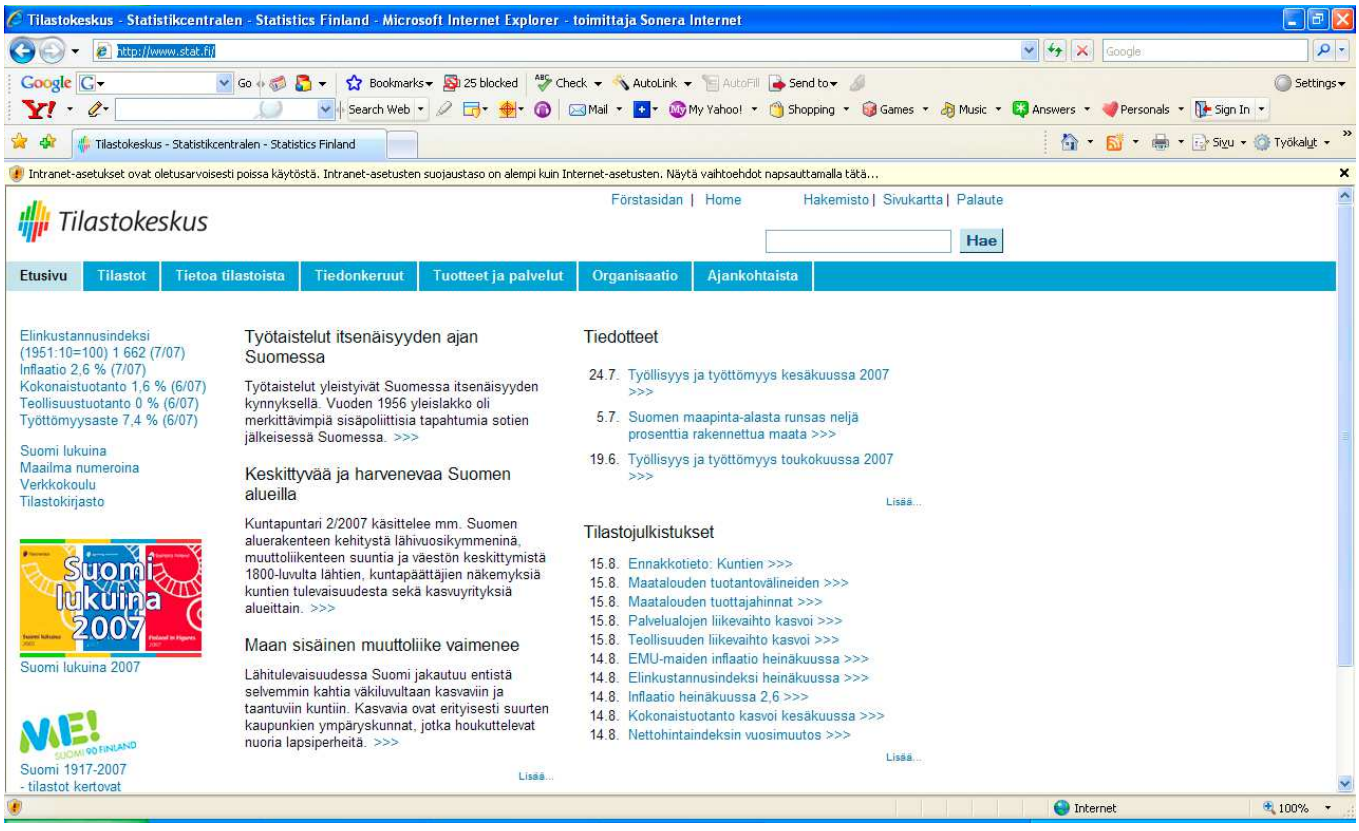
Symboli	Merkitys
”	Toisto
-	Ei yhtään
0	Suure pienempi kuin puolet käytetystä yksiköstä, vastaavasti merkitään 0,0; 0,00 jne.
.	Tieto epälooginen esitettäväksi
..	Tietoa ei ole saatu tai se on liian epävarma esitettäväksi
...	Salassapitosäännön alainen tieto
*	Ennakkotieto (suomalaisissa tilastoissa)

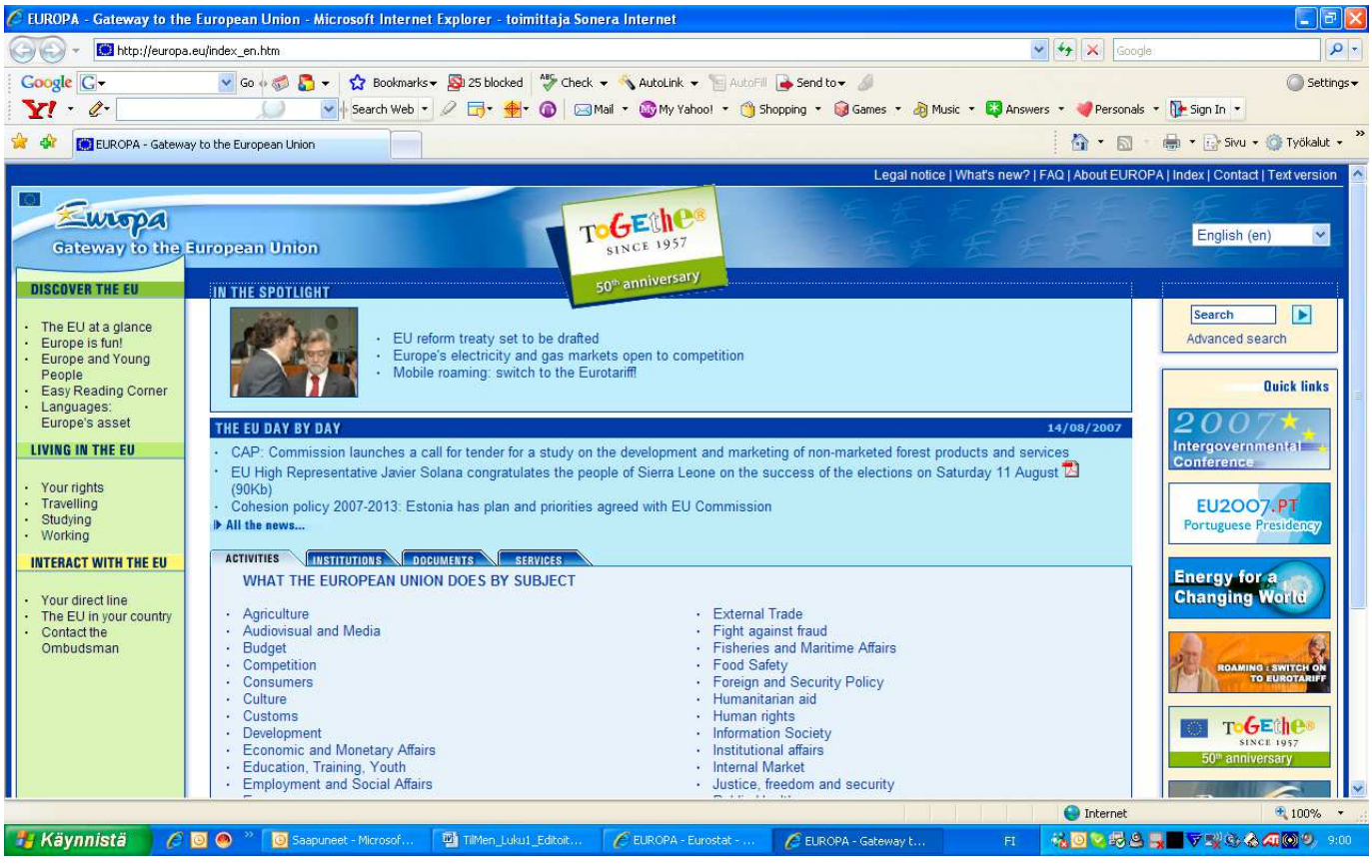
Huomautus
Suomessa käytetään desimaalierottimena pilkkua mutta monissa ulkomaisissa tilastoissa pistettä.

Internet

Internetin kautta on mahdollista päästä käsiksi **tietopankkeihin,** joiden **tietokannoista** voi poimia **hakukoneiden** tai **hakuohjelmistojen** avulla sopivat tiedot tai viitteet. Tietoa on tarjolla runsaasti, mutta sopivan tiedon löytäminen saattaa tuottaa pulmia. Sopivan tiedon löytämiseksi on kuitenkin kehitetty hakukoneita, joiden avulla löydetään tietopankki, josta puolestaan haettu tieto saattaa löytyä. Lisäksi tiedonhakuun on kehitetty **tiedon louhinta** (data mining). Ellei haettavaa tietoa löydy yhden hakukoneen tietokannoista, kannattaa aina jatkaa etsintää muiden hakukoneiden avulla.

Tilastoalan virtuaalikirjasto **WebStat** on Tilastokirjaston kokoama ja ylläpitämä tietokanta, johon on valittu internetin kautta vapaasti saatavilla olevia luotettavia virallisia tilastoja kaikkialta maailmasta. Tilastokeskuksen kotisivun osoite on http://www.stat.fi/. Tilastokeskuksen **StatFin**-tilastopalvelu tarjoaa Suomea koskevaa tilastotietoa internetin välityksellä. EU:n tilastoviraston, **Eurostatin,** kotisivun osoite on http://epp.eurostat.ec.europa.eu/.


Tilastokeskus - Statistikcentralen - Statistics Finland - Microsoft Internet Explorer - toimittaja Sonera Internet
Tilastokeskus
Förstasidan | Home
Hakemisto | Sivukartta | Palaute
Hae
Etusivu
Tilastot
Tietoa tilastoista
Tiedonkeruut
Tuotteet ja palvelut
Organisaatio
Ajankohtaista
Työtaistelut itsenäisyyden ajan Suomessa
Keskittyvää ja harvenevaa Suomen alueilla
Maan sisäinen muuttoliike vaimenee
Tiedotteet
Tilastojulkistukset
Suomi lukuina 2007
Suomi 1917-2007 - tilastot kertovat


EUROPA - Gateway to the European Union - Microsoft Internet Explorer - toimittaja Sonera Internet
Europa
Gateway to the European Union
Legal notice | What's new? | FAQ | About EUROPA | Index | Contact | Text version
English (en)
DISCOVER THE EU
LIVING IN THE EU
INTERACT WITH THE EU
IN THE SPOTLIGHT
THE EU DAY BY DAY
WHAT THE EUROPEAN UNION DOES BY SUBJECT
Search
Advanced search
Quick links

Internet-osoitteen perässä olevista kolmesta (tai kahdesta) kirjaimesta voi päätellä, millaisesta organisaatiosta on kyse.

Internet-tunnuksia

Tunnus	Merkitys
.com	Kaupallinen
.edu	Koulutus tai oppilaitos
.net	Verkko-operaattori
.org	Organisaatio, epäkaupallinen
.gov	Yhdysvaltojen valtion
.fi	Kansallinen tunnus (tässä suomalainen)

Tiedon omakohtainen kerääminen

Vain jos valmiita tilastoja ei ole, havaintoaineisto on kerättävä itse. Havaintoaineiston keräämistä otantamenetelmien avulla käsitellään luvussa 2.

1.6 Tilastollinen ajatteluprosessi

Tilastollisten menetelmien tunteminen on lähes kaikille kansalaisille tarpeellista, koska jo pelkästään sanoma- ja aikakauslehdissä ja televisiossa esitetään graafisia kuvioita ja tutkimustuloksia, joiden ymmärtäminen vaatii kuvanlukutaitoa, perustermien tuntemista ja tulosten arvioimiskykyä. Jos tutkimustuloksiin haluaa perehtyä perusteellisemmin tai tarkoituksena on itse laatia tutkimuksia, tilastollisten menetelmien hallinta on suorastaan välttämätöntä.

Tilastotiede (statistics) on menetelmätiede, jonka menetelmiä käytetään tutkimusten suunnittelussa ja toteutuksessa, tilastoaineiston kuvaamisessa ja analysoinnissa sekä ennusteiden laatimisessa ja johtopäätösten teossa.

Ilmiöiden kuvaamiseen ja käyttäytymisen ennakoimiseen käytetään usein mallia. Matemaattiset mallit voidaan jakaa **deterministisiin** ja **stokastisiin** malleihin. Kun ilmiötä kuvataan deterministisen mallin

avulla, tiettyjen alkuehtojen vallitessa voidaan määrittää tarkasteltavan ilmiön lopputulos. Esimerkkeinä tästä ovat monet fysiikan lait, kuten mekaniikan peruslait, Arkhimedeen laki ja Boylen laki. Stokastiset mallit perustuvat **todennäköisyyslaskentaan**. Stokastisia malleja käytetään, kun alkuehtojen perusteella ei voida varmasti määrittää tarkasteltavan ilmiön lopputulosta. Tällöin eri vaihtoehtoihin liittyvät tietyt esiintymistodennäköisyydet. Tällaisia ilmiöitä ovat esimerkiksi rahanheitto ja sään ennustaminen.

Kun tarkasteltavaa ilmiötä kuvataan stokastisen mallin avulla, joudutaan käyttämään tilastollisia menetelmiä. Vaikka laskenta hoidetaan tietokoneohjelmien avulla, tutkijan itsensä on huolehdittava tutkimusprosessin muista vaiheista. Kvantitatiivisen tutkimuksen tilastollisen analyysin periaate on seuraava:

Tilastollinen ajatteluprosessi.

TEHTÄVIÄ

1. Pohdi, onko intuitio tieteellinen menetelmä. Perustele vastauksesi.

2. Esitä esimerkki induktiivisesta ja deduktiivisesta päättelystä.

3. Millä mitta-asteikolla seuraavat muuttujat on mitattu:
- **a)** pituus
- **b)** paino
- **c)** älykkyysosamäärä
- **d)** kouluarvosana (1–5)
- **e)** postinumero
- **f)** kaupunginosa
- **g)** formulakilpailuiden pisteet (10, 8, 6, 5, 4, 3, 2, 1)
- **h)** puoluekanta
- **i)** silmien väri
- **j)** auton merkki
- **k)** henkilötunnus
- **l)** pankkitilin numero
- **m)** pankkitilin saldo
- **n)** pesuaineiden paremmuusjärjestys
- **o)** lämpötila kelvinasteina?

4. Pohdi,
- **a)** liittyykö mittauksen täydelliseen validiteettiin aina myös täydellinen reliabiliteetti, ja
- **b)** liittyykö mittauksen täydelliseen reliabiliteettiin vastaavasti aina myös täydellinen validiteetti.

5. Miksi tarvitaan luettelo tutkimusprosessin vaiheista?

6. Miksi tilastosymbolit on tunnettava?

7. Pohdi internetissä olevan tiedon luotettavuutta.

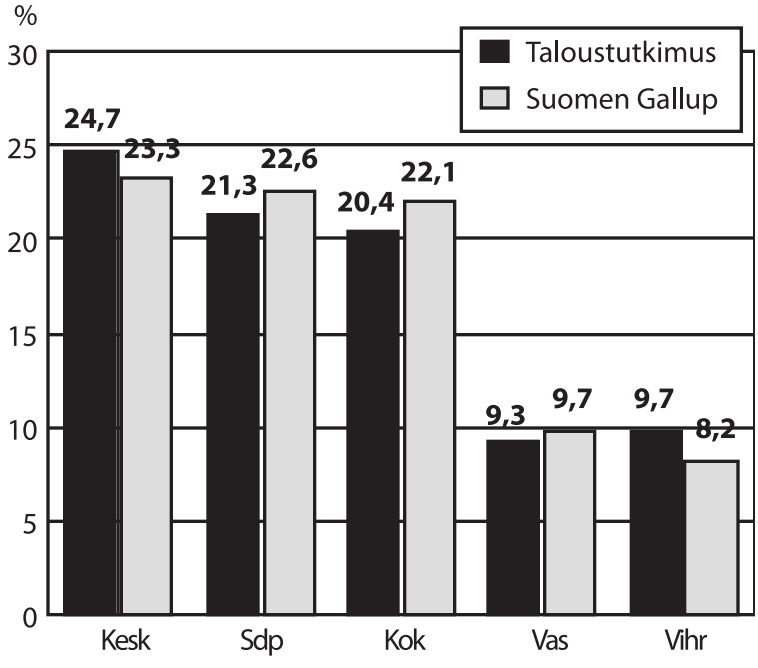
8. Vieraile Tilastokeskuksen kotisivulla. Lue opastussivut ja tee jokin tiedonhaku.

9. Mitä hyötyä on tilastollisen ajatteluprosessin tuntemisesta?

10. Pohdi, onko tutkimus eettistä ja pitäisikö sen olla sitä.

2
Tilastoaineiston kerääminen

Maaliskuussa 2007 julkaistiin Helsingin Sanomissa muutaman päivän sisällä kaksi puolueiden kannatusta mittaavaa tutkimustulosta. Toisen tutkimuksen oli tehnyt Yle Uutisten toimeksiannosta Taloustutkimus ja toisen HS-Gallupin toimeksiannosta Suomen Gallup. Molempien tutkimusten haastattelut oli tehty eduskuntavaalien alla maaliskuussa. Vastaukset voidaan esittää pylväsdiagrammina seuraavasti.



Suurimpien puolueiden kannatusarviot maaliskuussa 2007 Taloustutkimuksen ja Suomen Gallupin mukaan.

Molempien tutkimusten mukaan keskusta olisi Suomen suurin puolue, toisena olisi SDP ja kolmantena kokoomus. Lisäksi kävi ilmi, että vihreiden kannatus on Taloustutkimuksen mukaan neljäntenä 9,7 prosentin osuudella ja Suomen Gallupin mukaan viidentenä 8,2 prosentin osuudella.

Esiin nousee esimerkiksi seuraavia kysymyksiä: Onko puolueiden suuruusjärjestys sama kuin puolueiden lopullinen kannatus eduskuntavaaleissa? Kumpaa tutkimustulosta tulisi uskoa, eli onko toinen tutkimustulos ”oikea” ja toinen ”väärä”? Miten erot voidaan selittää? Näitä asioita selvitetään tämän luvun loppupuolella.

2.1 Kokonaistutkimus ja otantatutkimus

Edellisessä luvussa todettiin, että jos valmista tilastoaineistoa on käytettävissä, sitä kannattaa tietysti käyttää. Lisäksi kerrottiin, mistä valmiita tilastoja voi yrittää etsiä. Aina ei kuitenkaan ole saatavilla valmista tilastoaineistoa, jolloin se on kerättävä itse tai annettava sen hankkiminen jonkun muun tehtäväksi.

Tutkimus kiinnittää aina huomion johonkin tiettyyn kohteeseen, **perusjoukkoon** (population). Aluksi on ratkaistava kysymys, tehdäänkö tutkimus **kokonaistutkimuksena** (census), jossa perusjoukon jokainen **otantayksikkö** (sampling unit) otetaan tarkasteltavaksi. Toisena vaihtoehtona on tarkastella vain osaa perusjoukosta. Tällöin ajatellaan, että tämä osajoukko sopivasti valittuna voisi edustaa koko perusjoukkoa ja olisi ikään kuin perusjoukko pienoiskoossa. Tutkimalla tällaista osajoukkoa saadaan parhaimmillaan selville vastaavat ominaisuudet koko perusjoukosta. Tähän ideaan perustuvat **otantatutkimukset**.

Perusjoukon osajoukko, joka täyttää tietyt kriteerit, on nimeltään **otos** (sample). Tärkeimpänä ehtona on, että perusjoukon jokaisella otantayksiköllä on yhtä suuri mahdollisuus tulla valituksi otokseen. Ellei tämä ehto ole voimassa, osajoukkoa sanotaan **näytteeksi**. Perusjoukosta oikein poimittu otos on **edustava**, jos otokseen tulleiden otantayksiköiden ominaisuudet vastaavat perusjoukon vastaavia ominaisuuksia oikeassa suhteessa. Tällä tarkoitetaan sitä, että otos on kuin itse perusjoukko, mutta pienoiskoossa. Vaikka perusjoukosta poimittu otos olisi edustava, otoksesta lasketut tunnusluvut (ks. lukua 4) poikkeavat tavallisesti hieman perusjoukon tunnusluvuista. Tätä poikkeamaa sanotaan **otantavirheeksi** (sampling error).

Otantatutkimuksen vaiheet

Otantatutkimus etenee vaiheittain, ja jokaisessa vaiheessa kiinnitetään huomio johonkin tutkimuksen kannalta oleelliseen seikkaan. Kun vaiheet käydään läpi vuorotellen, saadaan otantatutkimus suoritetuksi. Otantatutkimuksen vaiheet ovat seuraavat:

- Tavoitteenmäärittely on keskeinen lähtökohta jokaisessa tutkimuksessa, myös otantatutkimuksessa.
- Perusjoukko määräytyy tavoitteenmäärittelyssä.
- Perusjoukko jaetaan otantayksiköihin. Otantayksikkö muodostuu yhdestä tai useammasta **tilastoyksiköstä**.
- Tavoitettavissa olevista otantayksiköistä muodostuu **kehikko** (frame).
- Valitaan otantamenetelmä.
- Määritellään otoskoko.
- Suunnitellaan otannan toteutus.
- Poimitaan otos.
- Käsitellään tilastoaineisto.
- Tulkitaan tulokset.
- Arvioidaan tuloksia kriittisesti.
- Kirjoitetaan raportti.

Otantatutkimus vai kokonaistutkimus?

Otantatutkimuksia tehdään paljon, koska niillä on useita hyviä puolia. Ne ovat myös tavallisesti paljon halvempia kuin kokonaistutkimukset. Lisäksi ne vievät vähemmän aikaa kuin kokonaistutkimukset, joten tulokset saadaan nopeammin käyttöön, ja niitä on helpompi hallita kuin kokonaistutkimusta. Joskus tosin kokonaistutkimusta on ainakin käytännössä mahdoton tehdä, erityisesti hyvin suuren perusjoukon osalta. Joskus taas kokonaistutkimusta ei ole edes mielekästä tehdä. Esimerkiksi Airam tuskin tekee kokonaistutkimusta valmistamiensa lamppujen kestoajan selvittämiseksi.

Kokonaistutkimus on yleensä järkevää tehdä, jos perusjoukko on pieni, esimerkiksi alle 100 otantayksikköä. Samoin tulee tehdä, jos otoskooksi tulisi yli puolet perusjoukon otantayksiköiden määrästä. Asianmukaisesti tehdyn otantatutkimuksen perusteella voidaan määrittää lopputuloksessa mahdollisesti esiintyvän virheen suuruus.

2.2 Otantamenetelmiä

Seuraavaksi pohditaan, miten otos pitäisi poimia, jotta se olisi edustava ja jotta siitä saadut tulokset voitaisiin yleistää koskemaan koko perusjoukkoa. Tätä varten on kehitetty otantamenetelmiä. Valitettavasti ei ole olemassa yhtä ja ainoaa otantamenetelmää, jota käyttämällä saataisiin aina paras mahdollinen tulos. Käytännön tilanteet poikkeavat toisistaan niin paljon, että eri tilanteisiin tarvitaan omat menetelmänsä.

Tässä luvussa esiteltävät otantamenetelmät voidaan jakaa kahteen ryhmään sen mukaan, onko otannan tuloksena otos vai näyte:

1. Todennäköisyyteen perustuvat menetelmät (tuloksena otos)
   - yksinkertainen satunnaisotanta
   - systemaattinen otanta
   - ositettu otanta
   - ryväsotanta
2. Harkintaan perustuvat menetelmät (tuloksena näyte)
   - harkinnanvarainen otanta
   - kiintiöotanta.

Yksinkertainen satunnaisotanta

Yksinkertainen satunnaisotanta (simple random sampling) eli lyhyesti **satunnaisotanta** on perusmenetelmä, jossa otantayksiköt poimitaan otokseen satunnaisesti. Tämä tarkoittaa sitä, että kukin otokseen poimittu otantayksikkö on tullut valituksi otokseen yhtä suurella todennäköisyydellä ja muista valinnoista riippumatta. Satunnaisotanta aloitetaan siten, että perusjoukon otantayksiköt numeroidaan ykkösestä eteenpäin. Tämän jälkeen päätetään otoskoko eli se, kuinka monta otantayksikköä otokseen poimitaan. Seuraavaksi kerätään eri satunnaislukuja niin monta kuin otoskoko edellyttää. Lopuksi perusjoukosta poimitaan satunnaislukujen osoittamat otantayksiköt.

Tarvittavat satunnaisluvut voidaan saada monella tavalla. Jokaisessa tietokoneessa on oma satunnaislukugeneraattori, ja useissa laskimissa on näppäin, jolla saadaan satunnaislukuja. Lisäksi on olemassa painettuja satunnaislukutauluja. Tietokoneen ja laskimen satunnaisluvut ovat tavallisesti 0:n ja 1:n välillä. Näitä lukuja sopivasti kertomalla saadaan tarvittavan suuria satunnaislukuja.

ESIMERKKI 1

Kaupungin yritysasiamiehen yritysluettelosta halutaan poimia 50 yritystä käsittävä yksinkertainen satunnaisotos. Tätä varten luettelon yritykset numeroidaan yhdestä eteenpäin, jolloin viimeinen yritys saa numerokseen 827. Seuraavaksi tarvitaan 50 satunnaislukua. Jos sama satunnaisluku esiintyy toistamiseen, se jätetään pois. Samoin menetellään, jos satunnaisluvuksi saadaan lukua 827 suurempi luku. Kun 50 eri satunnaislukua 1:n ja 827:n väliltä on saatu kerätyksi, yritysluettelosta poimitaan vastaavat yritykset. Näin satunnaisotos on saatu poimituksi.

Yksinkertaisen satunnaisotannan etuna on, että otos on varsin helppo poimia erityisesti silloin, kun saatavilla on valmiit luettelot eikä perusjoukko ole kovin suuri. Haittana taas voidaan pitää sitä, että heterogeenisen perusjoukon tapauksessa tulos saattaa vaihdella huomattavasti otoksesta toiseen. Lisäksi menetelmä on usein melko työläs, riittävän tarkkuuden saamiseksi otoskoko tulee suuremmaksi kuin useilla muilla menetelmillä eikä perusjoukosta mahdollisesti etukäteen tunnettuja tietoja voida käyttää hyväksi. Yksinkertaista satunnaisotosta ei aina voi edes poimia, jos kyse on esimerkiksi liukuhihnatuotannosta tai liikennevirrasta.

Systemaattinen otanta

Systemaattinen otanta (systematic sampling) on käyttökelpoinen otantamenetelmä, kun perusjoukko on mahdollista järjestää jonoon. Jonosta poimitaan tasaisin välimatkoin tai väliajoin otantayksiköt otokseen, kunnes lopullinen otos on kerätty. Systemaattinen otanta aloitetaan **poimintavälin** tai otoskoon määrittämisellä. Jos poimintaväliksi päätetään ottaa esimerkiksi 50, perusjoukosta valitaan otokseen joka viideskymmenes otantayksikkö. Tällöin otoskooksi tulee 2 % perusjoukon otantayksiköistä. Toisaalta otoskoko voidaan ilmoittaa prosentteina (p %) otantayksiköiden määrästä, jolloin poimintaväli $k = 100/p$ pyöristettynä kokonaisluvuksi.

Jos tunnetaan perusjoukon otantayksiköiden määrä N ja otoskoko n, poimintaväli k saadaan jakolaskun tuloksena: $k = \frac{N}{n}$. Laskettu poimintaväli pyöristetään lopuksi kokonaisluvuksi.

Ongelmana on vielä, mikä on ensimmäinen mukaan poimittava otantayksikkö. Tämä täytyy arpoa 1:n ja *k*:n väliltä esimerkiksi satunnaislukuja käyttämällä. Tämän jälkeen otokseen poimitaan tasavälisesti joka *k*:s otantayksikkö.

ESIMERKKI 2

Kaupungin yritysasiamiehen yritysluettelosta (ks. esimerkkiä 1) halutaan poimia systemaattisella otannalla otos siten, että otoskoko on 5 % yrityksistä. Poimintaväliksi saadaan

$k = \frac{100}{5} = 20$. Poimintaväliksi saadaan siten 20.

Ensimmäinen mukaan tuleva otantayksikkö saadaan poimimalla satunnaisluku 1:n ja 20:n väliltä. Tulokseksi saatiin luku 12. Mukaan tulevat siis numeroilla 12, 32, 52,... 812 olevat yritykset. Näitä on 41 kappaletta eli noin 5 % 827:stä niin kuin pitikin.

Systemaattinen otanta on kätevä otantamenetelmä silloin, kun saatavilla on valmis luettelo otantayksiköistä tai otantayksiköt voidaan järjestää jonoon. Se soveltuu siis hyvin esimerkiksi liikenteen, liukuhihnatuotannon, jäsenrekisterin, pääsylippua jonottavien tai kassajonossa seisovien otantayksiköiden poimimiseen otokseen. Systemaattinen otanta peittää perusjoukon tasaisesti. Huomattavasti aikaa, työtä ja kustannuksia säästyy, koska perusjoukon otantayksiköitä ei ole tarvinnut numeroida erikseen. Joskus numerointi ei olisi edes mahdollista.

Systemaattista otantaa on käytettävä varovasti, kun kyseessä ovat sykliset tilastoaineistot. Esimerkiksi liikennettä ei kannata valvoa vain iltapäivisin ruuhka-aikoina eikä kauppahallien kävijätutkimuksia tehdä vain aamupäivisin. Näissä tapauksissa tuloksissa esiintyy **harhaa** (bias) eli systemaattista virhettä, jolloin tunnusluvut ovat numeroarvoltaan systemaattisesti joko liian suuria tai liian pieniä.

Ositettu otanta

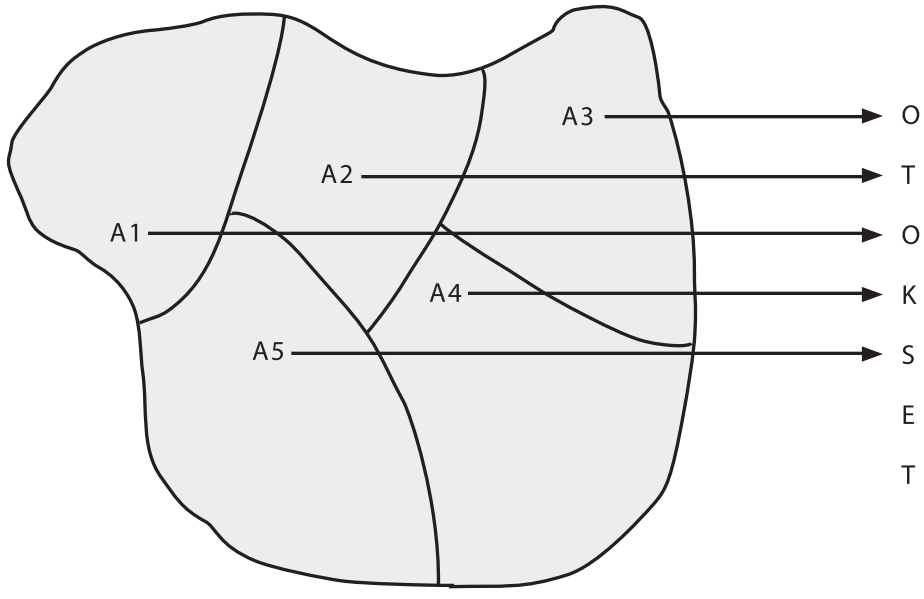
Ositetussa otannassa (stratified sampling) perusjoukko jaetaan osiin eli **ositteisiin** (strata). Perusjoukon ositus on tehtävä siten, että ositteet yhdessä kattavat koko perusjoukon mutta kukin otantayksikkö saa kuulua vain yhteen ositteeseen. Ositteiden pitäisi olla sisäisesti tutkittavan asian suhteen mahdollisimman homogeenisia. Ositteisiin

jakamisen jälkeen jokaisesta ositteesta poimitaan erikseen otos jotakin otantamenetelmää käyttäen. Ositteista poimitut otokset muodostavat yhdessä lopullisen otoksen.

Toisinaan täytyy pohtia, kuinka monta otantayksikköä poimitaan eri ositteista. Tarjolla on tällöin kolme vaihtoehtoa:

1. **tasainen kiintiöinti**, jossa jokaisesta ositteesta poimitaan sama määrä otantayksiköitä
2. **suhteellinen kiintiöinti**, jossa jokaisesta ositteesta poimitaan sama prosentuaalinen määrä otantayksiköitä
3. **paras kiintiöinti** (optimum allocation), jossa otetaan huomioon ositteiden yksilölliset piirteet, kuten koko, homogeenisuus ja otantakustannukset.

Käytännössä kuhunkin ositteeseen tulisi poimia vähintään 30 otantayksikköä.



Jokaisesta ositteesta poimitaan oma otos.

ESIMERKKI 3

Palataan vielä esimerkin 1 tilanteeseen, jossa yritysasiamiehen luetteloista halutaan poimia 80 yrityksen otos. Jos otos poimitaan ositetulla otannalla, ositteisiin jakaminen voidaan tehdä esimerkiksi henkilöstömäärän tai liikevaihdon mukaan. Tällöin pystytään takaamaan, että kaikenkokoiset yritykset ovat edustettuina lopullisessa otoksessa.

Jotta ositettua otantaa voitaisiin käyttää, perusjoukko on tunnettava etukäteen, sillä ellei perusjoukosta ole etukäteistietoja, ositteisiin jakaminen ei onnistu toivotulla tavalla. Jos perusjoukko voidaan jakaa järkevästi ositteisiin, ositettu otanta antaa yleensä tarkempia tuloksia kuin yksinkertainen satunnaisotanta. Ositetun otannan käyttö varmistaa myös erilaisten vähemmistöryhmien mukanaolon otoksessa.

Ryväsotanta

Ryväsotanta (cluster sampling) on kaksivaiheinen otantamenetelmä, jossa perusjoukko jaetaan aluksi toisensa poissulkeviin osajoukkoihin, **rypäisiin**. Jako tulisi tehdä siten, että kukin ryväs on ominaisuuksiltaan kuin koko perusjoukkoa. Rypäitä voisivat olla Suomessa esimerkiksi vaalipiirit, kaupunginosat tai ruokakunnat.

Varsinainen otanta tehdään siten, että rypäiden muodostamasta perusjoukosta poimitaan ensin satunnaisotos. Tämän jälkeen tehdään joko mukaan tulleiden rypäiden kokonaistutkimus tai saaduista rypäistä poimitaan vielä jokaisesta erikseen otos.

Ryväsotannan ero ositettuun otantaan on se, että kun ositetussa otannassa kukin osite on sisäisesti mahdollisimman homogeeninen ja ositteet siten poikkeavat keskenään selvästi toisistaan, ryväsotannassa kukin ryväs sisältää kaikki perusjoukon ominaisuudet ja ne ovat keskenään mahdollisimman samankaltaisia. Toisena erona on se, että ryväsotannassa vain jotkut rypäät tulevat poimituiksi mukaan ja siten tarkastelun kohteeksi, kun taas ositetussa otannassa kaikki ositteet ovat mukana lopullista otosta muodostettaessa.

ESIMERKKI 4

Lukiolaisten keskuudessa halutaan tehdä tutkimus, jossa selvitetään lukiolaisten suhtautuminen uuteen reaalikoejärjestelmään. Selvittäminen päätetään toteuttaa ryväsotannalla. Aluksi hankitaan taustatietoja: Suomessa on 428 lukiota, joissa on yhteensä noin 118 100 opiskelijaa. Kukin lukio muodostaa oman rypään, joten lukioiden muodostamasta perusjoukosta poimitaan satunnaisotannalla esimerkiksi 5 % eli 21 lukiota. Tämän jälkeen mukaan tulleista 21 lukiosta poimitaan jokaisesta systemaattisella otannalla tai ositetulla otannalla esimerkiksi jälleen 5 % opiskelijoista. Näin saatu otos edustaa todennäköisesti varsin hyvin koko lukiolaisten muodostamaa perusjoukkoa.

Ryväsotanta on varsin helppo tehdä, ja menetelmänä se on joustava ja edullinen. Sen heikkoutena pidetään kuitenkin sitä, ettei aina ole helppoa muodostaa rypäitä, jotka ovat toistensa kaltaisia. Myös tulosten tulkinta voi tuottaa vaikeuksia. Tulosten tarkkuus riippuu paljon siitä, kuinka hyvin rypäisiin jako onnistuu.

Harkinnanvarainen otanta

Harkinnanvaraisessa otannassa (convenience sampling) otantayksiköt poimitaan harkitusti, mutta kuitenkin siten, että pyritään mahdollisimman objektiiviseen ja tasapuoliseen tulokseen. Tällöinkään ei toteudu se otannan perusedellytys, että jokaisella otantayksiköllä on yhtä suuri mahdollisuus tulla valituksi otokseen. Tarkkaan ottaen tuloksena on siis näyte. Tämänkin menetelmän avulla voidaan silti saada melko luotettavia tuloksia, kunhan tutkimuksen tekoon kiinnitetään huomiota. Menetelmän käyttäminen vaatii siis aihealueen ja perusjoukon tuntemista, ja tulokset on tulkittava erityistä varovaisuutta noudattaen.

ESIMERKKI 5

Kun selvitetään kauppaketjun asiakastyytyväisyyttä, kyselyt on tarkoituksenmukaista tehdä jollakin yksittäisellä hetkellä kauppaketjun asiakkaiden joukossa, sillä kaikkia asiakkaita on mahdoton tavoittaa.

Harkinnanvaraisen otannan etuja ovat nopeus, joustavuus ja halpuus. Haittana taas saattaa olla se, ettei valittu osajoukko ehkä edustakaan koko perusjoukkoa, sillä otoksen valintamenettely ei takaa edustavuutta. Harkinnanvaraista otantaa käytetäänkin usein esitutkintaan, kyselylomakkeiden testaukseen, näkemyksen saamiseen sekä ideoiden ja hypoteesien kehittelyyn.

Kiintiöotanta

Kiintiöotannassa (quota sampling) perusjoukko tai kehikko jaetaan aluksi luokkiin (quota) joidenkin mitattavien ominaisuuksien osalta. Tällaisia ominaisuuksia voivat olla esimerkiksi sukupuoli, ikä, sosiaaliryhmä, kotipaikkakunta, lääni, puoluekanta tai uskonto. Seuraavaksi päätetään otoksen kokonaismäärä ja kuhunkin luokkaan mukaan otet-

tavien otantayksiköiden määrä. Itse otos poimitaan harkinnanvaraista otantaa tai systemaattista otantaa käyttämällä. Kun otanta etenee, kuhunkin osioon otetaan mukaan vain niin monta otantayksikköä kuin kiintiö vaatii. Kun sitten esimerkiksi miesten kiintiö on täynnä, miehiä ei enää voi poimia lisää otokseen. Otantaa jatketaan, kunnes jokainen kiintiö on täynnä.

Jotta kiintiöotantaa voidaan käyttää, tutkijalla pitää olla etukäteen tietoja perusjoukosta ja siitä, missä suhteessa otantayksiköt ovat jakautuneet eri luokkiin. Muutoin kiintiöiden muodostaminen on lähinnä arvailujen varassa ja tuloksiin saattaa tulla huomattavasti harhaa. Oikein tehtynä kiintiöotanta sen sijaan on nopea, joustava ja halpa menetelmä, ja sitä pidetään parempana kuin pelkkää harkinnanvaraista otantaa.

Harkinnanvaraisen otannan tavoin myös kiintiöotannassa tuloksena on tarkkaan ottaen näyte eikä otos.

ESIMERKKI 6

Eduskuntavaaleja varten halutaan selvittää puhelinhaastattelulla tehtävän kiintiöotannan avulla puolueiden kannatus. Tällöin on tärkeää, että otoksessa ovat oikeassa suhteessa muun muassa

- sukupuolijakauma
- ikäjakauma
- maantieteellinen jakauma
- jako kaupunkilaisiin ja maaseudulla asuviin
- jako sosiaaliryhmiin.

2.3 Otoskoon merkitys

Virhemarginaalit

On luontevaa ajatella, että samasta perusjoukosta saadaan otannan perusteella eri kerroilla hieman toisistaan poikkeavia tuloksia. Ratkaisevaa on tutkia, mitkä seikat vaikuttavat tulokseen ja sen tarkkuuteen. Vaikuttavia tekijöitä ovat sattuman lisäksi ainakin otantamenetelmä ja otoskoko.

Palataan luvun alussa esitettyyn tilanteeseen, jossa kaksi eri tutkimuslaitosta sai puolueiden kannatuksista erilaiset tulokset. Taloustutki-

mus sai kokoomuksen kannatukseksi 20,4 % ja Suomen Gallup 22,1 %. Vastaavasti Taloustutkimus sai SDP:n kannatukseksi 21,3 % ja Suomen Gallup 22,6 %. Voidaanko näiden tulosten perusteella sanoa, kumpi puolue on kannatukseltaan Suomen toiseksi suurin puolue?

Jokaiseen otantatutkimukseen sisältyy virhemahdollisuus. Virhemahdollisuuden riski on sitä suurempi, mitä pienemmästä otoksesta yritetään vetää johtopäätöksiä. Tämän vuoksi otoskoko yritetään saada mahdollisimman suureksi. Suuren otoksen poimimisen esteenä ovat usein kustannukset, aika ja hallittavuus. Sitä paitsi otoskoon kasvattaminen hyvin suureksi ei enää sanottavasti lisää tuloksen tarkkuutta. Seuraavassa taulukossa esitetään otoskoon vaikutus prosenttiosuuksien **virhemarginaaleihin** (margin of error). Taulukon luvut ovat prosenttiyksiköitä siten, että erehtymisriski on 5 %. Riskistä puhutaan tarkemmin luvussa 7.

Virhemarginaalitaulukko

	Otoskoko																				
%-luku	**30**	**50**	**75**	**100**	**125**	**150**	**200**	**250**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**	**1250**	**1500**	**2000**	**2500**	**3000**
1 tai 99	3,7	2,8	2,3	2,0	1,8	1,6	1,4	1,2	1,1	1,0	0,9	0,8	0,7	0,7	0,7	0,6	0,6	0,5	0,4	0,4	0,4
2 tai 98	5,2	4,0	3,2	2,8	2,5	2,3	2,0	1,7	1,6	1,4	1,2	1,1	1,0	1,0	0,9	0,9	0,8	0,7	0,6	0,5	0,5
3 tai 97	6,4	4,8	3,9	3,4	3,0	2,8	2,4	2,1	1,9	1,7	1,5	1,4	1,3	1,2	1,1	1,1	0,9	0,9	0,7	0,7	0,6
4 tai 96	7,3	5,6	4,5	3,9	3,5	3,2	2,7	2,4	2,2	1,9	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7
5 tai 95	8,1	6,2	5,0	4,3	3,9	3,5	3,0	2,7	2,5	2,1	1,9	1,7	1,6	1,5	1,4	1,4	1,2	1,1	1,0	0,9	0,8
6 tai 94	8,9	6,7	5,5	4,7	4,2	3,8	3,3	3,0	2,7	2,3	2,1	1,9	1,8	1,6	1,6	1,5	1,3	1,2	1,0	0,9	0,9
8 tai 92	10,1	7,7	6,2	5,4	4,8	4,4	3,8	3,4	3,1	2,7	2,4	2,2	2,0	1,9	1,8	1,7	1,5	1,4	1,2	1,1	1,0
10 tai 90	11,2	8,5	6,9	6,0	5,3	4,8	4,2	3,7	3,4	2,9	2,6	2,4	2,2	2,1	2,0	1,9	1,7	1,5	1,3	1,2	1,1
12 tai 88	12,1	9,2	7,5	6,4	5,8	5,2	4,5	4,0	3,7	3,2	2,9	2,6	2,4	2,3	2,1	2,0	1,8	1,6	1,4	1,3	1,2
15 tai 85	13,3	10,1	8,2	7,1	6,3	5,8	5,0	4,4	4,1	3,5	3,1	2,9	2,6	2,5	2,3	2,2	2,0	1,8	1,6	1,4	1,3
20 tai 80	14,9	11,4	9,2	7,9	7,1	6,5	5,6	5,0	4,5	3,9	3,5	3,2	3,0	2,8	2,6	2,5	2,2	2,0	1,8	1,6	1,4
25 tai 75	16,1	12,3	10,0	8,6	7,7	7,0	6,0	5,4	4,9	4,3	3,8	3,5	3,2	3,0	2,8	2,7	2,4	2,2	1,9	1,7	1,6
30 tai 70	17,1	13,0	10,5	9,1	8,1	7,4	6,4	5,7	5,2	4,5	4,0	3,7	3,4	3,2	3,0	2,8	2,5	2,3	2,0	1,8	1,6
35 tai 65	17,8	13,5	11,0	9,5	8,4	7,7	6,7	5,9	5,4	4,7	4,2	3,8	3,5	3,3	3,1	3,0	2,6	2,4	2,1	1,9	1,7
40 tai 60	18,3	13,9	11,3	9,7	8,7	7,9	6,8	6,1	5,6	4,8	4,3	3,9	3,6	3,4	3,2	3,0	2,7	2,5	2,1	1,9	1,8
45 tai 55	18,6	14,1	11,4	9,9	8,8	8,0	6,9	6,2	5,7	4,9	4,4	4,0	3,7	3,5	3,3	3,1	2,8	2,5	2,2	2,0	1,8
50	18,6	14,2	11,5	9,9	8,9	8,1	7,0	6,2	5,7	4,9	4,4	4,0	3,7	3,5	3,3	3,1	2,8	2,5	2,2	2,0	1,8

Virhemarginaalit (±) prosenttiyksikköinä, kun erehtymisriski on 5 %.

ESIMERKKI 7

Tämän luvun alussa mainituissa tutkimuksissa Taloustutkimus haastatteli lähes 2 000 ja Suomen Gallup 2 002 henkilöä. Suurimpien puolueiden osalta virhemarginaali on siis edellä olevan taulukon mukaan 1,8–1,9 prosenttiyksikköä kumpaankin suuntaan. Tämä tarkoittaa sitä, että Taloustutkimuksen saaman tuloksen mukaan arvio kokoomuksen kannatukselle on

(20,4 % – 1,8 pros.yks., 20,4 % + 1,8 pros.yks.) eli (18,6 %, 22,2 %).

Vastaavasti Suomen Gallupin saaman tuloksen mukaan arvio kokoomuksen kannatukselle on

(22,1 % – 1,8 pros.yks., 22,1 % + 1,8 pros.yks.) eli (20,3 %, 23,9 %).

Näihin tuloksiin sisältyy 5 %:n riski, että todellinen kannatus jää lasketun välin ulkopuolella. Tämän tarkempaa tulosta tutkimuksista ei voida saada. Johtopäätöksenä on, että kun virhemarginaalit otetaan huomioon, tutkimuslaitosten julkistamat tulokset eivät ole ristiriidassa keskenään kokoomuksen kannatuksen osalta.

SDP:n kannatuslukemat poikkeavat jonkin verran toisistaan. Kun otetaan huomioon kannatuksen virhemarginaali, arvio SDP:n kannatuksesta on Taloustutkimuksella 19,5 ja 23,1 %:n välillä, kun virhemarginaali on 1,8 prosenttiyksikköä. Suomen Gallupilla kannatus on puolestaan 20,7 ja 24,5 %:n välillä, kun virhemarginaali on 1,9 prosenttiyksikköä. Myöskään näissä tuloksissa ei ole keskenään mitään ristiriitaa.

Edellä olevia tuloksia tarkastelemalla ei kuitenkaan voida ratkaista, kumpi puolue – kokoomus vai SDP – lopulta on Suomen toiseksi suurin tai jopa suurin puolue.

Tarkastellaan vielä vihreiden kannatuslukemia. Taloustutkimus sai vihreiden kannatukseksi 9,7 % ja Suomen Gallup 8,2 %. Taulukosta käy ilmi, että Taloustutkimuksen saaman prosenttiluvun virhemarginaali on 1,3 prosenttiyksikköä ja Suomen Gallupin saaman prosenttiluvun virhemarginaali 1,2 prosenttiyksikköä. Taloustutkimuksen saaman tuloksen mukaan arvio vihreiden kannatuksesta on siis

(9,7 % – 1,3 pros.yks., 9,7 % + 1,3 pros.yks.) eli (8,4 %, 11,0 %).

Vastaavasti Suomen Gallupin saaman tuloksen mukaan arvio Vihreiden kannatukselle on

(8,2 % – 1,2 pros.yks., 8,2 % + 1,2 pros.yks.) eli (7,0 %, 9,4 %).

Nyt huomataan, ettei näissäkään tuloksissa ole keskenään mitään ristiriitaa. Molemmissa tapauksissa riski väärän johtopäätöksen tekemiseen on 5 %.

Jälkiviisaasti voidaan todeta luvun lopussa olevan tehtävän 15 perusteella, ettei kumpikaan tutkimuslaitos osannut ennakoida viiden tai edes kolmen suurimman puolueen keskinäistä suuruusjärjestystä. Edellä saatuja tuloksia katseltaessa huomataan myös, että oikeastaan on turhaa puhua kannatuksista yhden desimaalin tarkkuudella, kun prosenttilukujen kokonaisosatkaan eivät välttämättä ole oikein. Kun tämän pitää mielessään, osaa suhtautua oikein tutkimustuloksiin ja niistä tehtäviin johtopäätöksiin.

Otoskoko suhteellista osuutta määritettäessä

Sivun 38 taulukkoa voidaan käänteisesti käyttää myös otoskoon määrittämiseen, kun ensin päätetään virhemarginaali ja kun osataan arvioida suunnilleen tuloksen prosenttiluku. Käytännössä otoskooksi kannattaa ottaa jonkin verran taulukkoarvoa suurempi luku, sillä prosenttilukua ei tunneta tarkkaan etukäteen ja kato yleensä kuitenkin pienentää lopullista otoskokoa.

ESIMERKKI 8

Tuotteen *X* markkinaosuudeksi arvioidaan noin 20 %. Kuinka suuri otos on poimittava, jotta virhemarginaali olisi korkeintaan 2, 3, 4 tai 5 prosenttiyksikköä?

Sivun 38 taulukosta valitaan aluksi se rivi, jonka prosenttiluku on 20. Tämän jälkeen haetaan tältä riviltä luvut 2, 3, 4 ja 5. Ellei täsmälleen näitä lukuja löydy, valitaan lähinnä olevat luvut. Jotta virhemarginaali olisi vain kaksi prosenttiyksikköä, otoskoon on oltava 1 500. Jos tyydytään kolmen prosenttiyksikön virhemarginaaliin, otoskoon voi pienentää 700:aan. Neljän prosenttiyksikön virhemarginaali saadaan 400 otantayksikön otoksesta ja viiden prosenttiyksikön virhemarginaali 250 otantayksikön otoksesta. On kokonaan eri asia, tekeekö viiden prosenttiyksikön virhemarginaalilla enää mitään, sillä saatava arvio on jo hyvin väljä.

2.4 Otantatutkimuksen virhelähteitä

Kaikissa tutkimuksissa saattaa esiintyä eri lähteistä johtuvia virheitä. Tämä koskee myös otantatutkimuksia. Kun tiedostetaan virhelähteiden olemassaolo, virheitä on mahdollista yrittää välttää.

Otantatutkimuksissa kokonaisvirhe muodostuu otantavirheestä ja otantaan liittymättömistä virheistä. Otantavirhe voi syntyä esimerkiksi siksi, että valitaan tilanteeseen sopimaton otantamenetelmä, otos ei ole edustava tai otoskoko on liian pieni. Toisinaan otoksen **peitto** (coverage) ei ole hyvä. Tällöin otanta ei kata koko perusjoukkoa tai se kattaa perusjoukon ja vähän muutakin. Otantaan liittymättömiä virheitä ovat puolestaan vastaamiseen liittyvät virheet, vastaamattomuudesta johtuvat virheet ja tavoittamattomuudesta johtuvat virheet.

Vastaamiseen liittyviä virheitä ovat muun muassa vastaajasta johtuvat virheet, kuten tahallisten tai tahattomien väärien tietojen antaminen tai kysymysten väärinymmärtäminen. Vastaamiseen liittyviin virheisiin luetaan myös haastattelijasta johtuvat virheet, kuten vastausten väärin kirjaukset, vastaajien auttamiset ja annettujen ohjeiden noudattamattomuus. Vastaamiseen liittyviin virheisiin luetaan vielä tutkijasta johtuvat virheet, kuten huono tavoitteen määrittely, väärä kohderyhmä, puutteet tai epäselvyydet kyselylomakkeessa, puutteellinen tai väärä tilastoanalyysi sekä väärät tulkinnat ja johtopäätökset.

Vastaamattomuuden tai tavoittamattomuuden eli **kadon** (nonresponse) vuoksi lopullinen otoskoko poikkeaa tavoitteesta ja lopullisesta otoksesta saattaa jopa karsiutua pois tai olla aliedustettuna jokin osajoukko. Tällöin otoksesta tehtävät johtopäätökset saattavat muodostua virheellisiksi. Mahdollisen kadon vaikutuksen pienentämiseksi poimittavan otoskoon tulisi olla suunniteltua otoskokoa suurempi.

ESIMERKKI 9

Etukäteen arveltiin, että suomalaisista 90 % vastustaa rasismia. Tämän selvittämiseksi päätetään tehdä kyselytutkimus, jonka erehtymisriski on 5 % ja virhemarginaali 2 prosenttiyksikköä. Määritä otoskoko, jolla tulos voidaan selvittää, kun tällaisissa tutkimuksissa kato on noin 30 %.

Sivun 38 taulukon perusteella otoskoko ilman kadon vaikutusta on noin 900. Kadon huomioon ottava otoskoko on siten 900/0,70 = 1 286. Jotta kadon jälkeen riittävä otoskoko saavutettaisiin, lopulliseksi otoskooksi on valittava vähintään 1 300.

2.5 Kyselylomakkeet

Erityisesti haastattelututkimuksissa tiedot kerätään lomakkeilla: paperilomakkeilla, suoraan tietokoneavusteisesti tai internetin välityksellä. Lomakkeen käyttö nopeuttaa haastattelua, tarkentaa tietojen tallentamista, takaa tasapuolisuuden eri haastattelijoiden kesken ja nopeuttaa saatujen tulosten käsittelyä.

Kyselylomake on määrämuotoinen kysymyspatteristo, jolla kerätään informaatiota. Kyselylomakkeen käytöllä on useita tavoitteita. Tärkein niistä on muuntaa tutkijan tiedontarve kysymyksiksi, joihin vastaaja on kykenevä ja halukas vastaamaan. Jos lomakkeessa on valmiit vastausvaihtoehdot, vastauksista tulee niiden ansiosta myös yhdenmukaisia, mikä nopeuttaa lomakkeiden käsittelyä.

Lomakkeen tarkoituksena on myös minimoida vastaajan ponnistelut. Tämän tavoitteen saavuttamiseksi lomakkeessa tavallisesti esitetään valittaviksi ainakin osaan kysymyksistä jokin tai jotkin valmiit vaihtoehdot. Kyselylomakkeen tarkoituksena on lisäksi minimoida vastausvirheet, joita syntyy helposti, jos haastattelija yrittää tulkita vastaajan epätäsmällisiä tai epämääräisiä vastauksia. Kun tutkimus on päättynyt, lomakkeilla olevat tiedot voidaan säilyttää tulevia tarpeita varten.

Lomakkeiden laatiminen on vaativa tehtävä, joten on paikallaan esittää lomakkeen laatimista koskevia yleisiä ohjeita.

1. Määritellään, mitä tietoja halutaan kerätä.
2. Päätetään käytettävä menetelmä (esim. suora haastattelu, puhelinhaastattelu, internetkysely, kirjekysely).
3. Päätetään taustamuuttujat, joiden perusteella vastauksia voidaan myöhemmin jaotella.
4. Päätetään yksittäisten kysymysten sisältö.
5. Määritellään kysymysten rakenne.
   - Avointen kysymysten käsittely on hankalampaa kuin suljettujen kysymysten käsittely.
   - Monivalintakysymysten vaihtoehtojen on katettava kaikki mahdollisuudet, ja vaihtoehtojen täytyy olla toisensa poissulkevia.
6. Ratkaistaan asteikkojen valinta (esimerkiksi 1, 2, 3, 4, 5 tai –2, –1, 0, +1, +2).
7. Määritellään sanamuoto:
   - Kysymysten pitää olla lyhyitä ja yksikäsitteisiä.
   - Kysymykset eivät saa olla johdattelevia.
   - Yhdessä kysymyksessä selvitetään vain yhtä asiaa kerrallaan.
8. Tarkistetaan, ettei mitään tutkimuksen kannalta olennaista ole jäänyt pois.
9. Varmistetaan annettujen vastausten luotettavuus kontrollikysymyksillä.
10. Tarkistetaan, ettei kysymyksiä ole liian paljon, sillä liian pitkä kyselylomake heikentää vastausten laatua.
11. Määritellään kysymysten järjestys:
    - Helpot kysymykset sijoitetaan lomakkeen alkuun ja vaikeat ja arkaluontoiset lomakkeen loppuun.
12. Suunnitellaan lomakkeen ulkoasu (layout).
13. Testataan kyselylomake.

Kyselylomakkeen laatiminen on käytännössä jatkuva prosessi, jonka aikana lomaketta muokataan kerta toisensa jälkeen, kunnes lopulta mitään korjattavaa ei enää löydetä.

TEHTÄVIÄ

1. Voiko harkintaan perustuva näyte olla edustava?

2. Mikä ero on otantavirheellä ja harhalla?

3. Mikä ero on kehikolla ja perusjoukolla?

4. Mitä peitolla tarkoitetaan otantatutkimuksissa?

5. Miksi todennäköisyyteen perustuvia otantamenetelmiä suositaan enemmän kuin harkintaan perustuvia?

6. Muodosta 30 satunnaisluvun joukko 1:n ja 250:n väliltä.

7. Määritä poimintaväli systemaattisessa otannassa, kun
a) $n = 120$ ja $N = 7\,825$
b) $p = 4\,\%$.

8. Perusjoukko on jaettu ositteisiin A1–A5 siten, että eri ositteissa on seuraavat määrät otantayksiköitä: A1: 2 000, A2: 5 000, A3: 8 000, A4: 6 000, A5: 4 000. Kokonaisotosmääräksi valitaan 1 500 otantayksikköä. Kuinka monta otantayksikköä poimitaan eri ositteista, kun käytetään suhteellista kiintiöintiä?

9. Perusjoukko jaetaan osajoukkoihin ositetussa otannassa, ryväsotannassa ja kiintiöotannassa. Miten nämä menetelmät poikkeavat toisistaan osajoukkoihin jaettaessa?

10. Jos perusjoukon otantayksiköiden määrä kasvaa 100 000:sta 1 000 000:aan, onko otoskokoa nostettava samassa suhteessa yhtä tarkan tuloksen saamiseksi? Perustele vastauksesi.

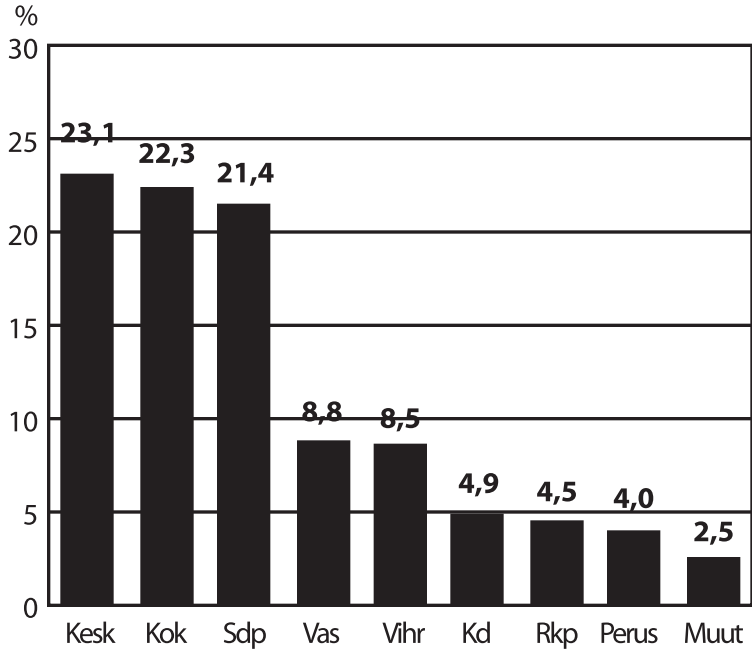
11. Tarkastellaan kahta tapausta. Tapaus 1: otoskoko on 500 ja kato 0 %. Tapaus 2: otoskoko on 1 000 ja kato 50 %. Saadaanko molemmista otoksista yhtä tarkat tulokset? Ellei näin ole, kumpaa tulosta mieluummin uskot ja miksi?

12. Arveltiin, että viallisten kappaleiden osuus tuotannosta on noin 4 %. Kuinka suuri otos tuotannosta on poimittava, jotta prosenttiluvun virhemarginaali olisi korkeintaan yksi prosenttiyksikkö, kun erehtymisriski on 5 %?

13. Puolueen P kannatukseksi on saatu 23,4 % haastattelemalla 1 968 henkilöä. Määritä virhemarginaali saadulle kannatukselle, kun erehtymisriskiksi valitaan 5 %.

14. Tarkastellaan vielä luvun alussa olleita mielipidemittauksia. Taloustutkimus sai RKP:lle 4,0 %:n kannatuksen ja Suomen Gallup 5,1 %:n kannatuksen. Määritä, millä välillä RKP:n kannatus tutkimuslaitosten saamien tulosten mukaan on, kun erehtymisriski on 5 %. Onko tuloksissa selvää ristiriitaa?

15. Kun vuoden 2007 eduskuntavaalien tulos vahvistettiin, eräs poliitikko sanoi, että erään tutkimuslaitoksen esittämät kannatusarviot olivat olleet ammattitaidottomasti tehtyjä. Saako väite tukea Taloustutkimuksen tai Suomen Gallupin tulosten osalta kolmen suurimman puolueen kannatuksesta, eli jääkö lopullinen tulos tutkimuslaitosten saamien tulosten virhemarginaalin sisään? Erehtymisriskinä käytetään tällaisissa tutkimuksissa tavanomaista viittä prosenttia. Tämän luvun alun kuviossa on esitetty Taloustutkimuksen ja Suomen Gallupin saamat kannatusarviot viidelle suurimmalle puolueelle. Puolueiden saama lopullinen kannatus eduskuntavaaleissa maaliskuussa 2007 on oheisen kuvion mukainen.



16. Kun vuoden 2007 eduskuntavaalien tulos vahvistettiin, eräs poliitikko sanoi: ”Perussuomalaisten kohdalla kaikki tutkimuslaitokset erehtyivät raskaasti.” Hän perusteli kantaansa sillä, että Taloustutkimus sai perussuomalaisten kannatukseksi 2,5 % ja Suomen Gallup 2,0 %. Saako väite tukea Taloustutkimuksen tai Suomen Gallupin tulosten osalta, eli jääkö lopullinen tulos tutkimuslaitosten saamien tulosten virhemarginaalin sisään? Erehtymisriskinä käytetään tällaisissa tutkimuksissa tavanomaista viittä prosenttia. Puolueiden saama lopullinen kannatus vuoden 2007 eduskuntavaaleissa on esitetty tehtävässä 15.

17. Esitä otantamenetelmien jaottelu kaaviokuvana.

18. Muodosta otantatutkimuksen virhelähteitä esittävä kaaviokuva.

19. Mitä taustamuuttujia kyselylomakkeissa yleensä käytetään? Entä mihin niitä käytetään?

20. Miksi kyselylomake pitää testata etukäteen?

3
Tilastojen tutkiminen

Kuvailevassa tilastotieteessä (descriptive statistics) tutkimuksen kohteena olevasta ilmiöstä kerätyt tiedot esitetään tiivistetysti **taulukoina**, **graafisina kuvioina** ja **tilastollisina tunnuslukuina**. Tässä luvussa perehdytään taulukoiden muodostamiseen ja tilastojen graafiseen esittämiseen.

Taulukoiden ja kuvioiden tarkoituksena on

- tiivistää suuren tilastoaineiston sanoma asiayhteyden kannalta keskeisiin seikkoihin
- osoittaa totuudenmukaisesti tilastotietojen sisältämä fakta
- houkutella lukijaa ajattelemaan asiaa
- korostaa, että taulukon tai kuvion sanoma syntyy vertailusta
- elävöittää tilastoraporttia.

3.1 Luokittelu ja taulukointi

Havaintomatriisi

Tilastoaineiston käsittely aloitetaan muodostamalla kerätystä tilastotiedosta aluksi **havaintomatriisi**.

Tilasto-yksikkö	**Muuttuja** **1**	**2**	**…**	***p***
1	x_{11}	x_{12}		x_{1p}
2	x_{21}	x_{22}		x_{2p}
3	x_{31}	x_{32}		x_{3p}
.	.	.		.
.	.	.		.
.	.	.		.
n	x_{n1}	x_{n2}		x_{np}

Havaintomatriisi muodostuu tilastoyksiköihin liittyvien muuttujien arvoista seuraavasti:

– Kukin vaakarivi sisältää yhteen tilastoyksikköön liittyvät muuttujien arvot.
– Jokainen pystyrivi eli sarake sisältää yhden muuttujan kaikki arvot.

Pystyriviä sanotaan suoraksi **jakaumaksi** tai **jakaumavektoriksi**.

ESIMERKKI 1

Eräässä 50 tilastoyksikön otoksessa tutkittiin muuttujia

x_1 = sukupuoli (1 = nainen, 2 = mies)
x_2 = ikä (vuosia)
x_3 = bruttopalkka (€/kk)
x_4 = asumiskustannukset (€/kk)
x_5 = asumismuoto (1 = omakotitalo, 2 = pari- tai rivitalo, 3 = kerrostalo, 4 = muu).

Tilastoaineistosta koottiin havaintomatriisi.

Tilasto-yksikkö	Muuttuja 1	2	3	4	5
1	2	35	3 950	575	3
2	1	47	2 700	690	2
3	1	28	3 950	380	1
4	2	33	2 550	465	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
50	2	19	2 100	400	3

Havaintomatriisista voidaan muun muassa lukea, että otoksen neljäs tilastoyksikkö on 33-vuotias omakotitalossa asuva mies, jonka kuukausitulot ovat 2 550 euroa ja asumiskustannukset 465 euroa kuukaudessa.

Tilastoaineiston luokitteluun ja taulukointiin tutustutaan seuraavan esimerkin avulla.

ESIMERKKI 2

Päivittäistavaroita myyvä liike teki asiakastutkimuksen poimimalla 120 henkilön otoksen asiakasrekisteristään. Otokseen valituille lähetettiin kirjallinen kyselylomake. Lomakkeen palautti 102 henkilöä. Tutkittavia muuttujia olivat

- sukupuoli (1 = nainen, 2 = mies)
- ikä (vuosina)
- kotitalouden (perheen) koko (henkilöä)
- asuminen (1 = omistusasunto, 2 = vuokra-asunto, 3 = työsuhdeasunto)
- koulutus (1 = peruskoulututkinto, 2 = ammatillisen toisen asteen tutkinto, 3 = ylioppilastutkinto, 4 = ammattikorkea-koulututkinto, 5 = yliopistotutkinto)
- kotitalouden nettotulot (€/kk)
- kotitalouden päivittäistavaroiden ostot (€/kk: 1 = alle 250, 2 = 250–499, 3 = 500–749, 4 = 750–999, 5 = 1 000 tai enemmän)
- tyytyväisyys liikkeen tuotevalikoimaan (asteikko 1–5: 1 = erittäin tyytymätön, 5 = erittäin tyytyväinen)
- tyytyväisyys liikkeen hintatasoon (asteikko 1–5: 1 = erittäin tyytymätön, 5 = erittäin tyytyväinen).

Muuttujien arvot kerättiin havaintomatriisiksi ASIAKAS. Tietokoneelle sen voi poimia osoitteesta www.sanomapro.fi > Oppilaitoksille > Ammatti- ja korkeakoulut > Ammattiverkko.

Luokittele ja taulukoi muuttujat
a) koulutus,
b) kotitalouden nettotulot.

a) Muuttuja ”koulutus” saa ”arvot” 1–5. Havaintomatriisin kyseisestä sarakkeesta voidaan laskea, kuinka monta kutakin arvoa on. Laskeminen kannattaa antaa tietokoneohjelman tehtäväksi. Kaikissa tilasto-ohjelmissa on komennot, joilla laskeminen voidaan suorittaa. Saatua taulukkoa sanotaan **frekvenssitaulukoksi**. Havaintoarvojen lukumääriä puolestaan sanotaan **frekvensseiksi** (merkitään tavallisesti *f*).

Tilasto-ohjelman tulostama frekvenssitaulukko on seuraava:

Frequency Distribution – Quantitative		
Data	**Frequency**	**Percent**
1	16	15,7
2	22	21,6
3	10	9,8
4	33	32,4
5	21	20,6
	102	100,0

Frekvenssitaulukossa frekvenssit on esitetty myös prosentteina kokonaisfrekvenssistä 102, jolloin on saatu suhteelliset frekvenssit.

Jos lasketaan suhteellisten frekvenssien summa, saadaan 100,1 %. Taulukkoon merkitään kuitenkin kokonaisfrekvenssiksi aina 100 %. Prosentin kymmenesosan ero johtuu pyöristysvirheistä.

b) Taulukoitaessa nettotuloja ei voida menetellä kuten edellä, koska erilaisia havaintoarvoja on lähes 100. Tällöin muuttujien arvot on luokiteltava siten, että lähellä toisiaan olevat arvot sijoitetaan samaan luokkaan. Luokittelu aloitetaan etsimällä muuttujan suurin ja pienin arvo:

Suurin arvo	5 100 €
Pienin arvo	940 €
Erotus	4 160 €

Tässä tapauksessa sopiva luokkien lukumäärä voisi olla kuusi. Lasketaan suurimman ja pienimmän havainnon erotus jaettuna luokkien lukumäärällä eli 4 160 € /6 ≈ 693 €. Pyöristetään saatua tulosta sopivasti ylöspäin, esimerkiksi arvoon 700 €.

Valitaan ensimmäisen luokan alarajaksi jokin sopiva arvo pienimmän havaitun muuttujan arvon (940 €) alapuolelta, esimerkiksi 900 €. Koska havaintoarvot ovat 10 €:n tarkkuudella, ensimmäiseen luokkaan sijoittuvat arvot, jotka kuuluvat 900 ja 1 590 €:n välille, toiseen luokkaan arvot, jotka kuuluvat 1 600 ja 2 290 €:n välille jne.

Koska jokaisen luokan **leveys** eli **peitto** on sama (700 €), luokitusta sanotaan **tasaväliseksi**.

Syötetään tietokoneohjelmaan tarvittavat arvot:
1) ensimmäisen luokan alaraja 900
2) luokkaväli 700.

Frekvenssitaulukko on seuraava:

Frequency Distribution – Quantitative								
tulot							*cumulative*	
lower		*upper*	*midpoint*	*width*	*frequency*	*percent*	*frequency*	*percent*
900	<	1 600	1 250	700	14	14,0	14	14,0
1 600	<	2 300	1 950	700	26	26,0	40	40,0
2 300	<	3 000	2 650	700	28	28,0	68	68,0
3 000	<	3 700	3 350	700	19	19,0	87	87,0
3 700	<	4 400	4 050	700	6	6,0	93	93,0
4 400	≤	5 100	4 750	700	7	7,0	100	100,0
					100	100,0		
2 missing or invalid data values								

Muuttujan arvoja on vain 100, koska kaksi vastaajaa ei ollut ilmoittanut kotitaloutensa nettotuloja. Ohjelma ilmoittaa myös, kun arvoja puuttuu.

Taulukkoon tulostuvat myös **kumulatiiviset** eli **kertymäfrekvenssit**. Kumulatiivinen frekvenssi (merkitään tavallisesti *F*) osoittaa niiden havaintoarvojen määrän, jotka ovat enintään yhtä suuria kuin tarkasteltavan luokan yläraja. Jotta kumulatiivisia frekvenssejä voidaan käyttää, tutkittava muuttuja pitää olla mitattu vähintään **järjestysasteikolla**.

Edellisessä taulukossa kumulatiivinen frekvenssi 87,0 % osoittaa, että otokseen joutuneista kotitalouksista 87 prosentilla nettotulot olivat alle 3 700 euroa kuukaudessa. Ohjelma tulostaa automaattisesti myös **luokkakeskukset**, jotka saadaan luokkien ala- ja ylärajojen keskiarvona.

Seuraavat tilastoaineiston luokittelua ja taulukointia varten annetut ohjeet ovat viitteellisiä. Niitä kannattaa noudattaa mahdollisuuksien mukaan.

Luokitteluohjeita

1. Sopiva luokkien lukumäärä on 4–10, ja siihen vaikuttavat muun muassa tilastoaineiston laajuus ja maksimi- ja minimiarvon erotus.
2. Luokittelu tulisi pyrkiä tekemään tasaväliseksi.
3. Avoimia luokkia, joista on ilmoitettu vain ala- tai yläraja, tulisi välttää, koska se vaikeuttaa muun muassa eräiden tunnuslukujen laskemista ja graafista esitystä.
4. Luokkaväliksi ja luokkien alarajoiksi (joskus ylärajoiksi) valitaan sopivan pyöreitä lukuja.

Esimerkiksi ohjeita 2 ja 3 ei kannata noudattaa, jos muutama havaintoarvo poikkeaa huomattavasti valtaosasta muita arvoja. Jos poikkeavia arvoja on vain yksi, se käytännössä usein poistetaan havaintoarvojen joukosta, jolloin ongelma häviää. Jos poikkeavia arvoja on muutamia, ensimmäinen tai viimeinen luokka (tai molemmat) merkitään avoimeksi.

Ristiintaulukointi

Muuttujien taulukoinnin tarkoitus on antaa yleiskuva jakaumasta tai muuttujien välisestä **riippuvuudesta**. Muuttujien välistä riippuvuutta tutkitaan **ristiintaulukoimalla**, jolloin saadaan **kaksiulotteinen frekvenssijakauma**.

ESIMERKKI 3

Kun tutkitaan alustavasti havaintomatriisia ASIAKAS, havaitaan, että muuttujien 5 (koulutus) ja 8 (tyytyväisyys liikkeen tuotevalikoimaan) välillä on ilmeisesti riippuvuutta siten, että koulutustason kasvaessa tyytyväisyys liikkeen tuotevalikoimaan vähenee. Asian selvittämiseksi muuttujat ristiintaulukoidaan.

Crosstabulation

		Tuoteval					
		1	**2**	**3**	**4**	**5**	**Total**
Koulutus	**1**			1	5	10	16
	2			6	7	9	22
	3			2	3	5	10
	4		10	15	5	3	33
	5	2	11	6	2		21
	Total	2	21	30	22	27	102

Taulukko osoittaa, että arvelu muuttujien välisestä riippuvuudesta on luultavasti oikea. Pelkän peruskoulun käyneistä kukaan ei ollut kovin tyytymätön liikkeen tuotevalikoimaan, vaan lähes kaikki heistä olivat melko tai erittäin tyytyväisiä siihen. Sen sijaan akateemisesti koulutettujen vastaukset painottuivat arvoihin 2 ja 3 eli kukaan heistä ei ollut erittäin tyytyväinen valikoimaan.

Riippuvuus voidaan varmistaa tilastollisen testien avulla. Niitä käsitellään luvussa 7.

3.2 Tilastojen graafinen esittäminen

Taulukoiden lisäksi tilastoja havainnollistetaan erilaisilla graafisilla esityksillä eli **diagrammeilla**. Diagrammi on taulukkoa parempi esitystapa, koska se havainnollistaa esitettävän asian yhdellä silmäyksellä ja tuo esiin myös poikkeamat. Lisäksi kuvio kiinnittää taulukkoa paremmin huomiota ja tuo tilastoraporttiin elävyyttä.

Diagrammit laaditaan käytännössä aina tietokoneohjelmien avulla. Tilasto-ohjelmien lisäksi kaikilla taulukkolaskentaohjelmilla voidaan tuottaa tasokasta grafiikkaa. Esimerkiksi Excelin avulla voidaan laatia kymmeniä erilaisia kuvioita. Peruskuviotyypit ovat

1. **pylväsdiagrammit**
2. **viivadiagrammit**
3. **sektoridiagrammit** ja **muut pinta-aladiagrammit**.

Jokaisesta kuviotyypistä on useita erilaisia variaatioita.

Kuvioita piirrettäessä on otettava huomioon eräitä periaatteita. Tärkein niistä on, että kuvion tulee antaa lukijalle yhdellä silmäyksellä oikea käsitys tilastosta. Kun tilastoja esitetään kuvioina, lukijalle syntyy käsitys esitettävästä asiasta yleensä kuvion pinta-alojen perusteella. Pinta-alojen tulee siis olla suoraan verrannollisia esitettävien luokkien kokoon. Tilastokeskus käyttää tästä nimitystä **visuaalisen vertailtavuuden periaate**.

Pylväsdiagrammit

Peruspylväskuvion avulla kuvataan yleensä epäjatkuvan muuttujan arvojen jakaumaa. Ryhmiteltyjä pylväitä käytetään, kun esitetään samassa kuviossa usean muuttujan arvojen jakaumat. Pylväät piirretään usein vaakasuoraan (palkkikuvio), kun kuvataan nominaaliasteikon muuttujaa eli muuttuja on kvalitatiivinen (ks. lukua 1). Ositetut pylväät muodostuvat siten, että osien summana saadaan kokonaisuus. Histogrammia käytetään kuvaamaan luokitellun jatkuvan muuttujan arvojen jakaumaa.

Seuraavissa esimerkeissä tutustutaan erilaisiin pylväsdiagrammeihin.

ESIMERKKI 4

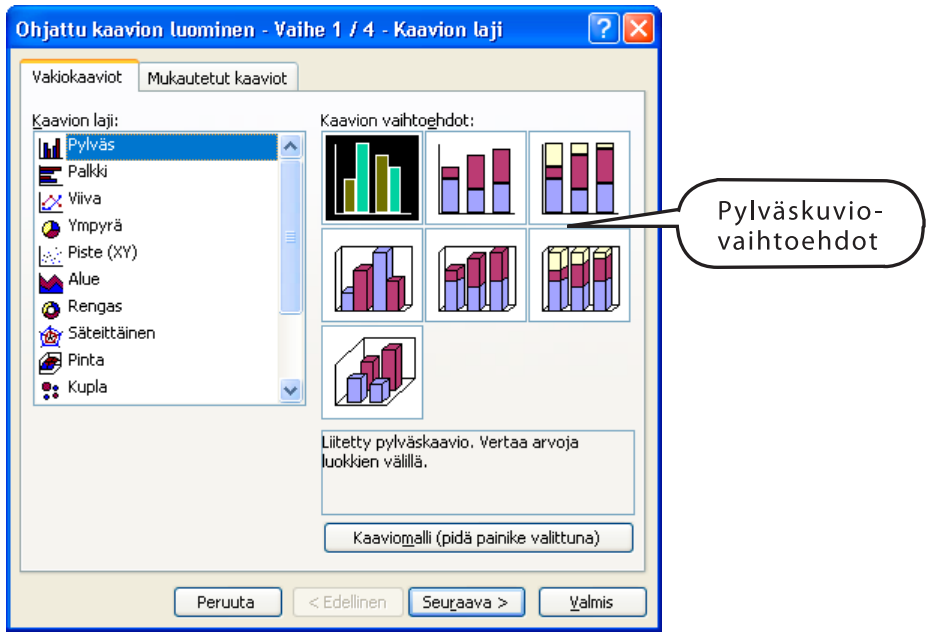
Kuntien ja kuntayhtymien sosiaali- ja terveystoimen toimintamenot sekä kuntien rahoitus kyseiseen toimintaan (mrd. €) vuosina 2000–2007 olivat seuraavat (lähde: Kuntaliitto):

Vuosi	Toimintamenot	Kuntien rahoitus
2000	11,3	7,2
2001	12,2	7,8
2002	12,9	8,1
2003	13,6	8,3
2004	14,1	8,4
2005	15,2	8,8
2006	16,1	9,3
2007	17,0	9,8

a) Esitä kuntien ja kuntayhtymien sosiaali- ja terveystoimen toimintamenojen kehitys pylväsdiagrammina.
b) Esitä samassa kuviossa sekä toimintamenojen että kuntien niihin osoittaman rahoituksen kehitys.
c) Mitä päätelmiä voit tehdä taulukon ja kuvioiden perusteella?

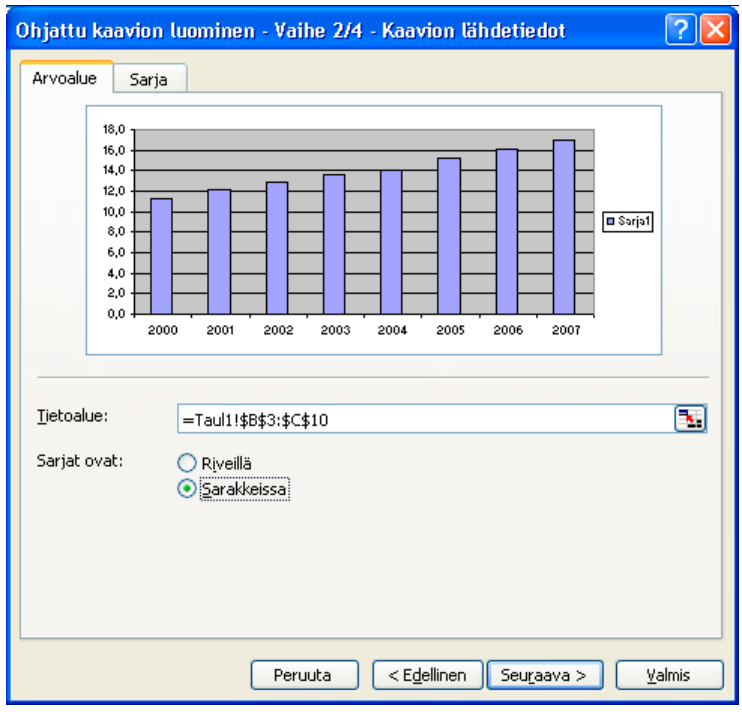
Tehdään kuvio Excelin ohjattua kaavion luomistoimintoa käyttämällä.

Määritetään aluksi kuvion (kaavion) laji, joka tässä tapauksessa on pylväsdiagrammi.

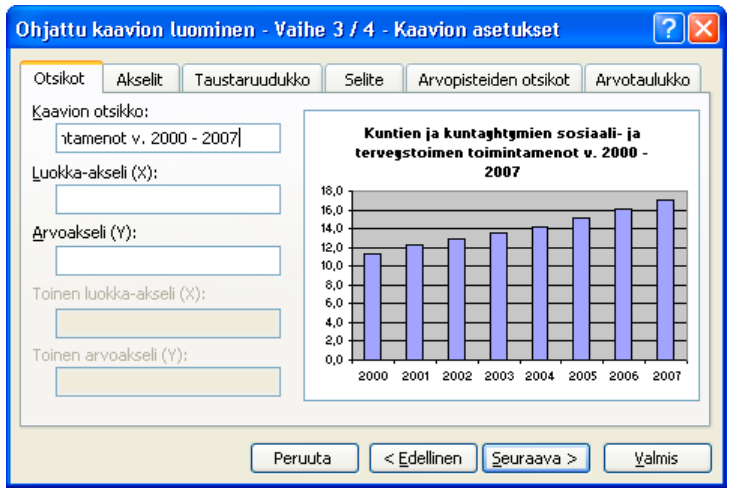


Valitaan sen jälkeen jokin esillä olevista vaihtoehdoista.

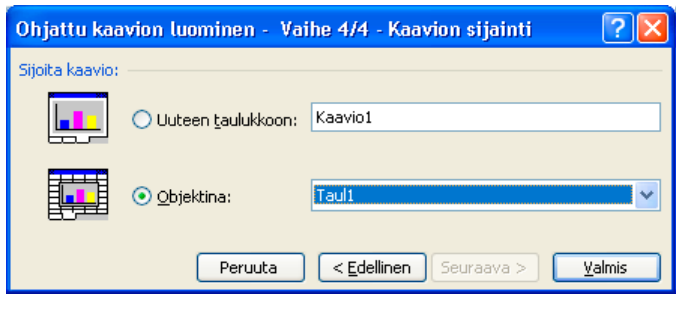
Toisessa vaiheessa määritetään solut, joissa kuvion piirtämistä varten tarpeelliset lähdetiedot ovat.



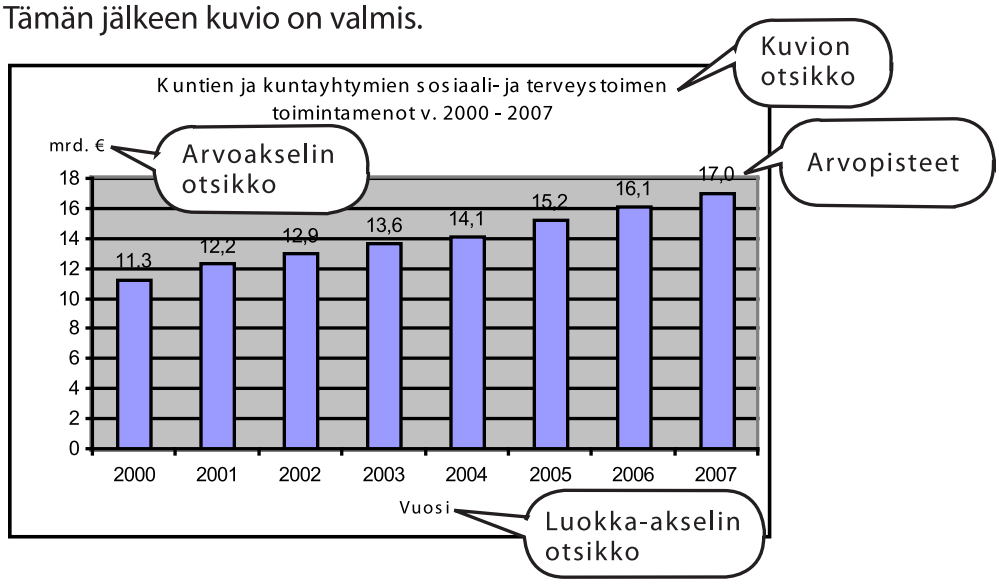
Määritetään seuraavaksi kuvion asetukset (otsikot, selitteet, arvopisteet, asteikkojen mittayksiköt, apuviivat ym.).



Neljännessä vaiheessa on tarpeen mukaan mahdollista määrittää kuvion sijainti.

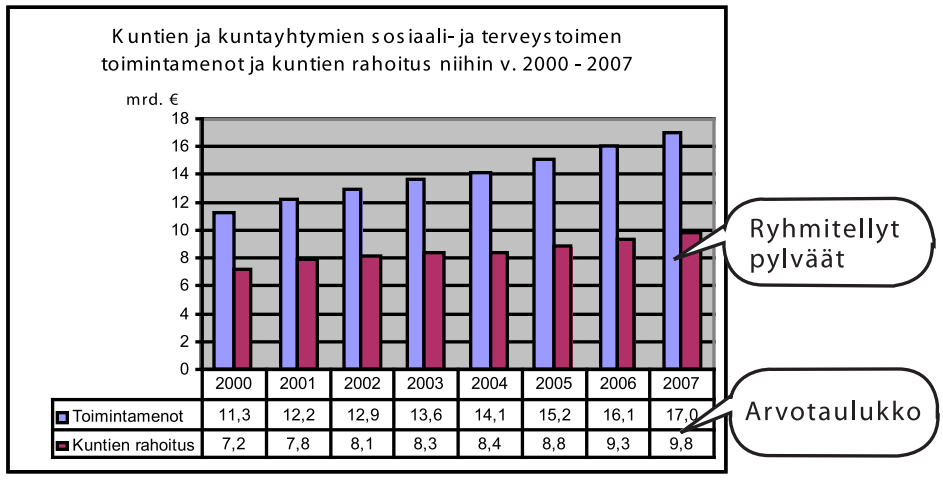


Tämän jälkeen kuvio on valmis.



Kuvio on peruspylväsdiagrammi. Pylväät on piirretty pystyasentoon. Lisäksi kuvioon on lisätty otsikot sekä toimintamenojen määrä (arvopisteet) eri vuosina pylväiden yläosaan.

b) Piirretään pylväät siten, että toimintamenot ja kuntien niihin osoittama rahoitus ovat vierekkäin (**ryhmitellyt pylväät**).



Tässä muuttujien arvot on merkitty kuvion alapuolelle. Jos ne merkittäisiin pylväiden päihin, kuviosta tulisi sotkuinen.

c) Heti käy ilmi, että kuviosta on selvästi helpompi kuin taulukosta yhdellä silmäyksellä havaita toimintamenojen ja kuntien rahoituksen kehitys. Molemmat ovat kasvaneet verraten ripeästi, ja viimeisten vuosien aikana kasvu on kiihtynyt. Esimerkiksi kuntien rahoitus on vuosina 2005–2007 kasvanut 500 miljoonaa euroa vuodessa. Ei olekaan ihme, että monien kuntien taloudellinen tilanne on vuosikymmenen aikana huonontunut ja monet kunnat ovat säästöjen etsimisen lisäksi joutuneet korottamaan veroprosenttiaan.

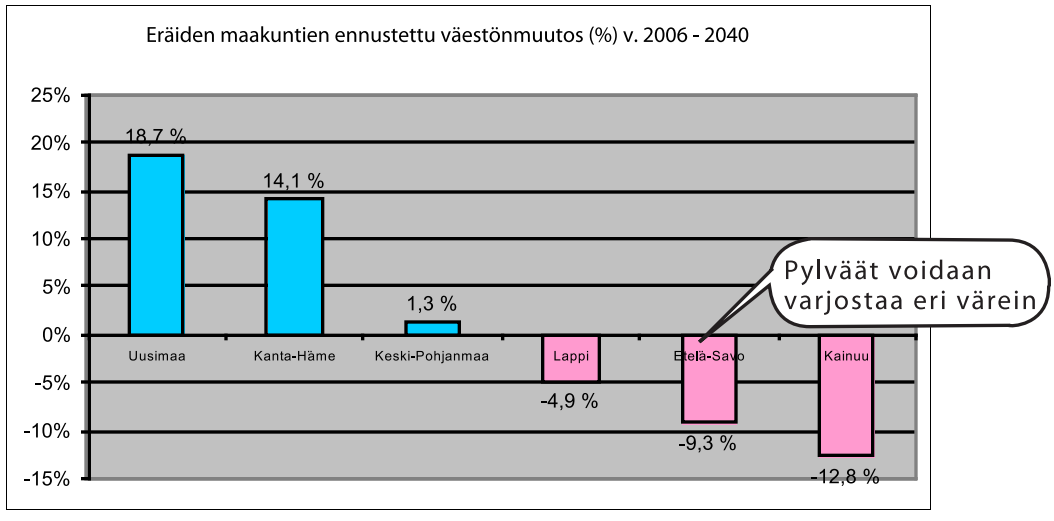
Pylväskuvioilla voidaan esittää myös tilastoja, joissa osa muuttujan arvoista on positiivisia, osa negatiivisia.

ESIMERKKI 5

Tilastokeskuksen laatiman ennusteen mukaan väestönmuutos (%) eräissä maakunnissa vuodesta 2006 vuoteen 2040 on seuraava:

Maakunta	Ennustettu väestönmuutos
Uusimaa	18,7 %
Kanta-Häme	14,1 %
Keski-Pohjanmaa	1,3 %
Lappi	–4,9 %
Etelä-Savo	–9,3 %
Kainuu	–12,8 %

Esitä tilasto pylväskuviona. Mitä voi päätellä Suomen väestön sijoittumisesta maakuntiin vuonna 2040, jos ennuste toteutuu?



Kuviossa positiiviset ja negatiiviset muutokset on esitetty erivärisinä. Esitys antaa viitteitä siitä, että Suomen väestö on keskittymässä Etelä- ja Länsi-Suomeen ja että Itä- ja Pohjois-Suomea uhkaa väestökato.

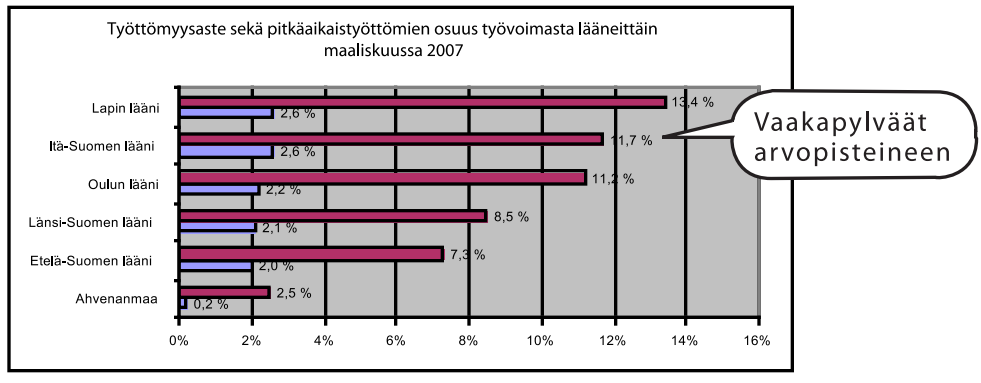
Pylväskuvio voidaan piirtää myös siten, että pylväät ovat vaakasuorassa.

ESIMERKKI 6

Työttömyysaste sekä pitkäaikaistyöttömien osuus työvoimasta lääneittäin maaliskuussa 2007 oli seuraava (lähde: Työministeriö):

Lääni	Työttömyysaste (%)	Pitkäaikaistyöttömien osuus työvoimasta (%)
Lapin lääni	13,4	2,6
Itä-Suomen lääni	11,7	2,6
Oulun lääni	11,2	2,2
Länsi-Suomen lääni	8,5	2,1
Etelä-Suomen lääni	7,3	2,0
Ahvenanmaa	2,5	0,2

Esitä tilasto pylväskuviona siten, että pylväät on piirretty vaakasuoraan (**palkkikuvio**). Miten tilastoa voidaan tulkita?



Kuviosta käy ilmi, että työttömyysaste alenee selvästi etelään ja länteen päin mentäessä. Toisaalta pitkäaikaistyöttömien osuus vaihtelee paljon vähemmän kuin työttömyysaste. Yleensä pitkäaikaistyöttömät ovat monin tavoin syrjäytyneitä eivätkä he yleensä ole työmarkkinakelpoisia. Heidän osuutensa Manner-Suomessa on runsaat 2 prosenttia työvoimasta. Ahvenanmaa poikkeaa muusta maasta sekä työttömyysasteen että pitkäaikaistyöttömien osalta selvästi. Ahvenanmaalla on käytännössä täystyöllisyys eikä syrjäytyneitäkään juuri ole.

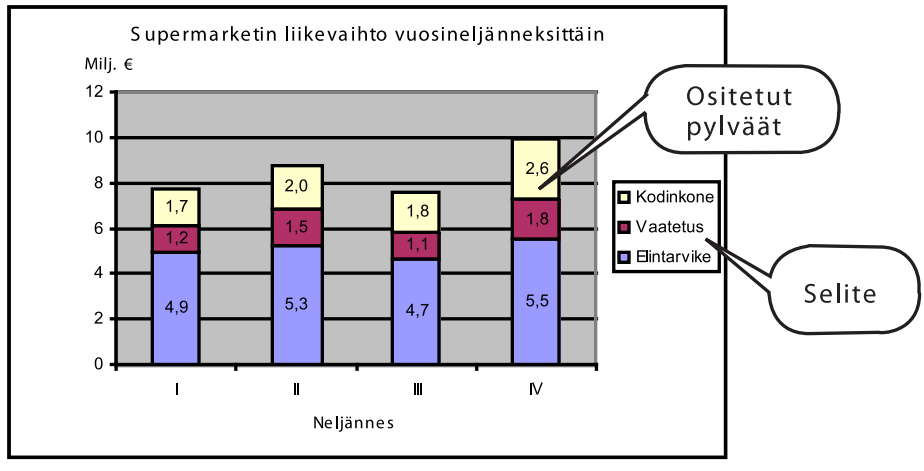
ESIMERKKI 7

Supermarketissa on elintarvikeosasto, vaatetusosasto ja kodinkoneosasto. Osastojen liikevaihdot olivat erään vuoden eri neljänneksillä seuraavat:

Vuosi-neljännes	Elintarvikeosasto (milj. €)	Vaatetusosasto (milj. €)	Kodinkoneosasto (milj. €)
I	4,9	1,2	1,7
II	5,3	1,5	2,0
III	4,7	1,1	1,8
IV	5,5	1,8	2,6

Esitä Supermarketin liikevaihto pylväskuviona vuosineljänneksittäin siten, että siitä käy ilmi jokaisen vuosineljänneksen myynti osastoittain.

Tässä tapauksessa kuvio kannattaa piirtää siten, että osastojen myyntiä kuvaavat pylväät ovat päällekkäin (**ositetut pylväät**), jolloin päällekkäin olevien pylväiden summa osoittaa osastojen yhteisen myynnin.



Kuvion oikealle puolelle on liitetty selite, josta ilmenee, millä tavalla kunkin osaston myyntiä esittävä pylväs on varjostettu. Kuviosta havaitaan kokonaisliikevaihdon muutokset eri neljänneksinä mutta myös osastojen liikevaihtojen muutokset.

Erikoistapaus pylväskuviosta on **histogrammi**. Sen avulla kuvataan jatkuvan luokitellun muuttujan jakaumaa, jolloin pylväät piirretään yhteen ilman välejä.

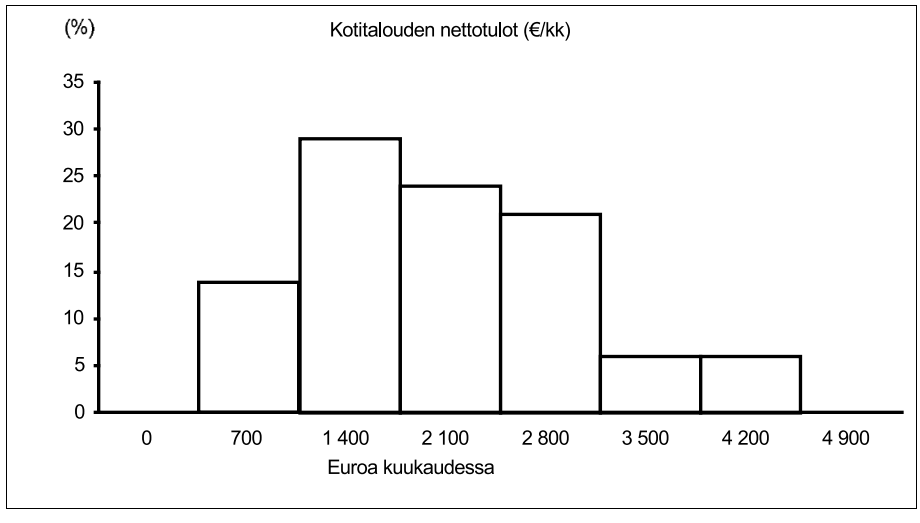
ESIMERKKI 8

Kohdan 3.1 esimerkissä 2 luokiteltiin havaintomatriisin ASIAKAS muuttuja 6 (kotitalouden nettotulot). Jakauma luokiteltuna ja sitä vastaava histogrammi ovat seuraavat:

Frequency Distribution – Quantitative

tulot							*cumulative*	
lower		*upper*	*midpoint*	*width*	*frequency*	*percent*	*frequency*	*percent*
900	<	1 600	1 250	700	14	14,0	14	14,0
1 600	<	2 300	1 950	700	26	26,0	40	40,0
2 300	<	3 000	2 650	700	28	28,0	68	68,0
3 000	<	3 700	3 350	700	19	19,0	87	87,0
3 700	<	4 400	4 050	700	6	6,0	93	93,0
4 400	≤	5 100	4 750	700	7	7,0	100	100,0
					100	100,0		

2 missing or invalid data values



Viivadiagrammit

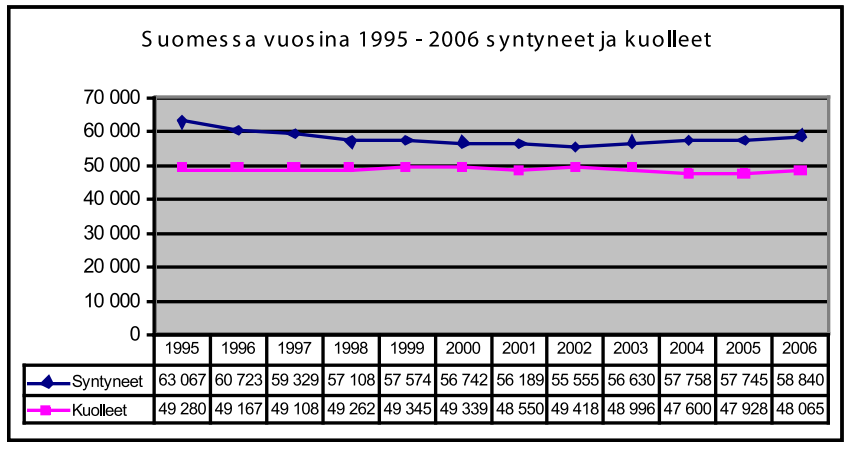
Viivadiagrammia käytetään ennen kaikkea ajan mukana muuttuvien suureiden eli **aikasarjojen graafiseen esittämiseen**. Viivadiagrammien tarkoitus on antaa yleiskuva tutkittavan suureen **suunnasta** ja **muutosnopeudesta** eri aikoina. Pystyakselin katkaisu on viivadiagrammia piirrettäessä yleensä mahdollista.

ESIMERKKI 9

Suomessa vuosina 1995–2006 syntyneiden ja kuolleiden määrät on esitetty seuraavassa tilastossa (lähde Tilastokeskus).

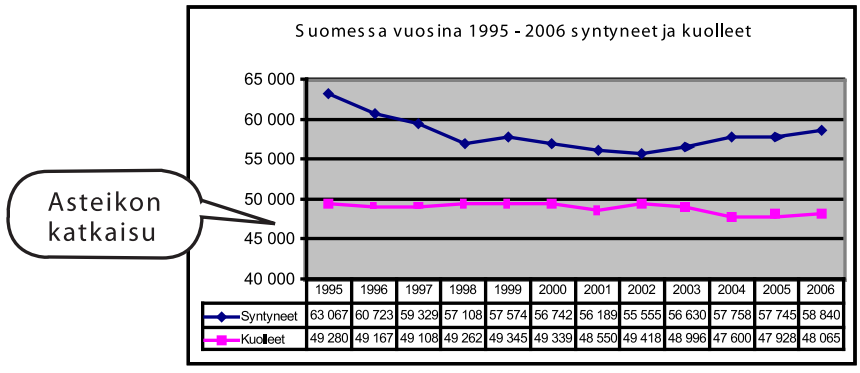
Vuosi	Syntyneet	Kuolleet
1995	63 067	49 280
1996	60 723	49 167
1997	59 329	49 108
1998	57 108	49 262
1999	57 574	49 345
2000	56 742	49 339
2001	56 189	48 550
2002	55 555	49 418
2003	56 630	48 996
2004	57 758	47 600
2005	57 745	47 928
2006	58 840	48 065

Esitä syntyneiden ja kuolleiden määrät samassa kuviossa viivadiagrammina.



Kuvio ei ole kovin havainnollinen, koska asteikko alkaa nollasta. Tällöin syntyneitä ja kuolleita esittävät murtoviivat ovat kuvion ylälaidassa eikä vaihtelusta saa kovinkaan havainnollista kuvaa.

Piirretään diagrammi siten, että pystyakseli aloitetaan arvosta 40 000.



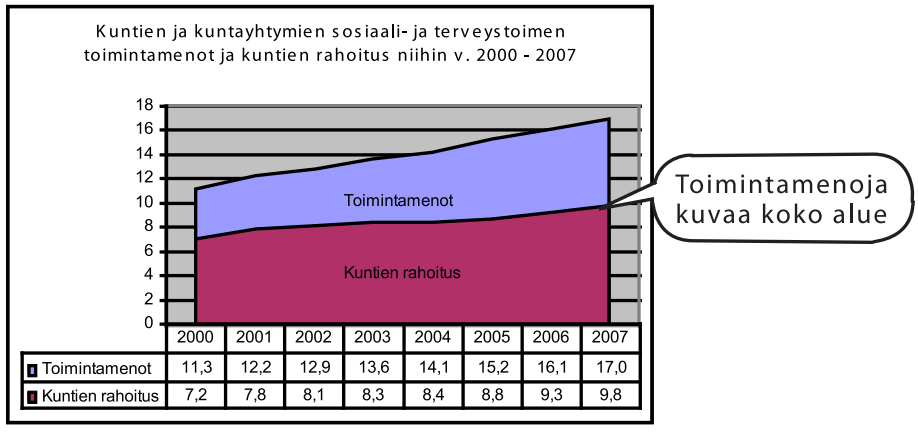
Murtoviivojen väli osoittaa syntyneiden ja kuolleiden erotuksen. Havaitaan, että vuodesta 1995 vuoteen 2002 syntyvyys on yhtä poikkeusta lukuun ottamatta pienentynyt. Kuolleisuus on samalla aikavälillä pysynyt lähes ennallaan, joten syntyneiden ja kuolleiden erotus on pienentynyt. Vuoden 2002 jälkeen syntyvyys on kääntynyt nousuun ja kuolleisuus laskuun, joten erotus on kasvanut.

Mieti, missä asioissa voidaan hyödyntää edellisen tilaston tietoja.

Erikoistapaus viivadiagrammista on **aluekuvio**. Aluekuviossa koko viivan tai viivojen alapuolelle jäävä alue väritetään tai rasteroidaan.

ESIMERKKI 10

Esitä esimerkin 4 b) tilasto (Kuntien ja kuntayhtymien sosiaali- ja terveystoimen toimintamenot ja kuntien niihin osoittama rahoitus) aluekuviona.



Kuviosta ilmenee pylväskuviota selvemmin kuntien rahoituksen suhde toimintamenoihin ja molempien kehitys, mutta yksittäisen vuoden arvoja on tällaisesta kuviosta vaikeampi lukea kuin pylväskuviosta.

Sektoridiagrammit

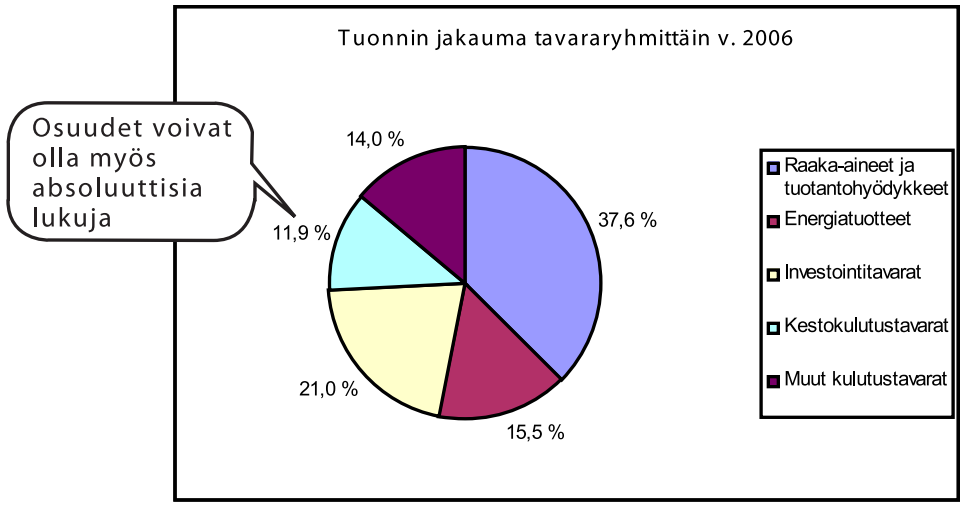
Sektoridiagrammeja (pie, ”piirakka”) käytetään, kun halutaan **havainnollistaa kokonaisuuden jakautumista osiin**. Jokaisen luokan kokoa edustaa sektorin pinta-ala, joka on suoraan verrannollinen luokan kokoon. Tavallisesti vertaillaan suhteellisia osuuksia (prosenttiosuuksia).

ESIMERKKI 11

Tavaratuonnin suhteellinen jakauma (%) Suomeen käyttötarkoituksen mukaan vuonna 2006 oli seuraava (lähde: Tullihallitus):

Tavaratuonti	%
Raaka-aineet ja tuotantohyödykkeet	37,6
Energiatuotteet	15,5
Investointitavarat	21,0
Kestokulutustavarat	11,9
Muut kulutustavarat	14,0

Esitä tavaratuonnin jakauma sektoridiagrammina.



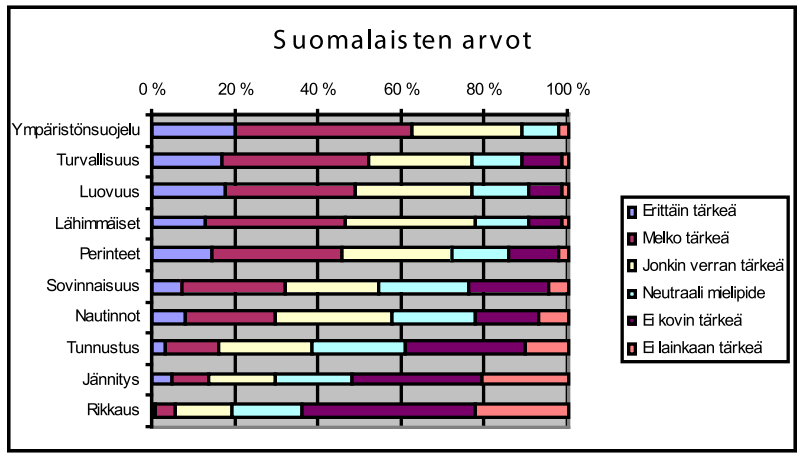
Sektoridiagrammin lisäksi kokonaisuuden jakautumista osiin voidaan kuvata palkkikuvion avulla. Palkkikuviota käytetään varsinkin silloin, kun verrataan keskenään useita muuttujia, joiden arvot ovat osa kokonaisuudesta.

ESIMERKKI 12

Suomalaisten arvoja mittaavassa tutkimuksessa saatiin seuraavat tulokset (lähde: World Values Survey):

Arvo	Erittäin tärkeä	Melko tärkeä	Jonkin verran tärkeä	Neutraali mielipide	Ei kovin tärkeä	Ei lainkaan tärkeä
Ympäristönsuojelu	20 %	43 %	26 %	9 %	0 %	2 %
Turvallisuus	17 %	35 %	25 %	12 %	10 %	1 %
Luovuus	18 %	31 %	28 %	14 %	8 %	1 %
Lähimmäiset	13 %	34 %	31 %	13 %	8 %	1 %
Perinteet	15 %	31 %	26 %	14 %	12 %	2 %
Sovinnaisuus	7 %	25 %	23 %	21 %	19 %	5 %
Nautinnot	8 %	22 %	28 %	20 %	15 %	7 %
Tunnustus	3 %	13 %	23 %	22 %	29 %	10 %
Jännitys	5 %	9 %	16 %	18 %	31 %	21 %
Rikkaus	1 %	5 %	13 %	17 %	42 %	22 %

Esitä arvojen jakauma graafisesti samassa kuviossa.



Kuviosta käy ilmi, että ympäristönsuojelu on noussut suomalaisten mielissä keskeiseksi asiaksi. Sitä pitää erittäin tai melko tärkeänä yli 60 % suomalaisista. Myös turvallisuus, lähimmäiset, perinteet ja luovuus ovat kansalaisten mielestä tärkeitä. Sen sijaan rikkautta piti erittäin tai melko tärkeänä selvästi alle 10 % suomalaisista.

Pinta-aladiagrammien yksi erikoistapaus on **teemakartta**, jonka avulla voidaan kuvata alueellisia tietoja. Tässä yhteydessä sitä ei kuitenkaan käsitellä.

Muita kuvioita

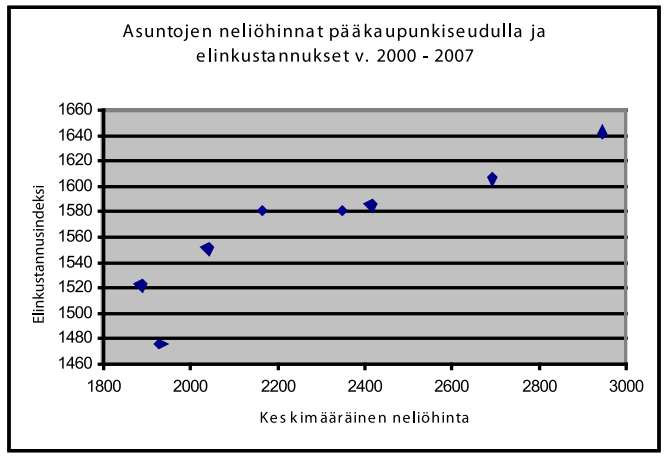
Muista kuvioista keskeisiä ovat **parvi-** tai **hajontakuviot**, joiden avulla tutkitaan ja havainnollistetaan muun muassa kahden muuttujan välistä riippuvuutta. Kuvion x-akselille merkitään havaintoparin toisen muuttujan arvot ja y-akselille toisen muuttujan arvot.

ESIMERKKI 13

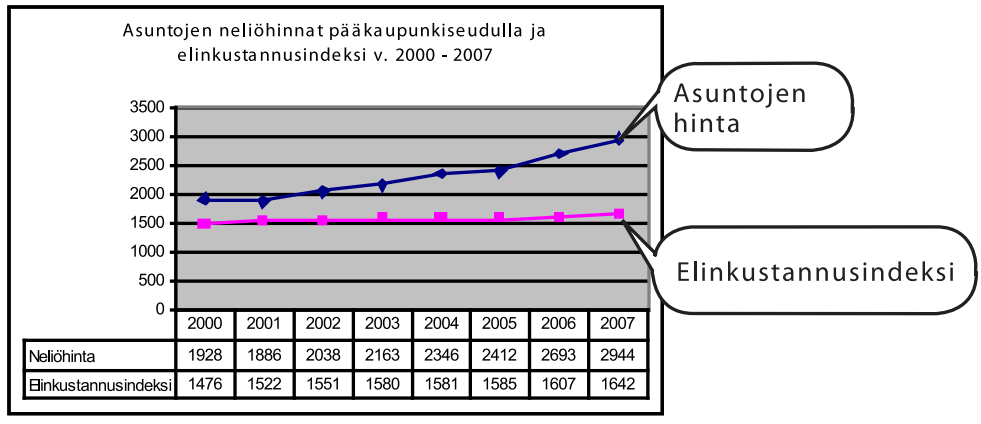
Tutkimuksessa haluttiin selvittää pääkaupunkiseudun asuntojen hintakehitystä verrattuna muiden elinkustannusten kallistumiseen. Asian selvittämiseksi verrattiin vanhojen kerrostalohuoneistojen keskimääräisiä neliöhintoja (euroa/m^2) elinkustannusindeksin pistelukuun vuosina 2000–2007. Keskimääräiset neliöhinnat ja elinkustannusindeksin pisteluku olivat vuosina 2000–2007 seuraavat (lähde: Tilastokeskus):

Vuosi	Neliöhinta	Elinkustannusindeksi
2000	1 928	1476
2001	1 886	1522
2002	2 038	1551
2003	2 163	1580
2004	2 346	1581
2005	2 412	1585
2006	2 693	1607
2007	2 944	1642

Piirrä parvikuvio. Mitä johtopäätöksiä voit taulukon ja kuvion perusteella tehdä?



Heti voidaan havaita, että neliöhintojen ja elinkustannusindeksin pisteluvun välillä on selvä yhteys: molemmat keskimäärin kasvavat vuosien kuluessa. Tämä on tietenkin luonnollista, koska asuntojen hintojen kehitys ei voi olla kokonaan erisuuntaista kuin muiden hintojen kehitys. Taulukkoa tarkastelemalla havaitaan kuitenkin myös, ettei neliöhintojen ja elinkustannusindeksin pisteluvun välinen yhteys ole lineaarinen. Kun piirretään viivadiagrammi, josta ilmenee molempien muuttujien kehitys, voidaan havaita, että asuntojen hinnat ovat nousseet elinkustannuksia nopeammin.



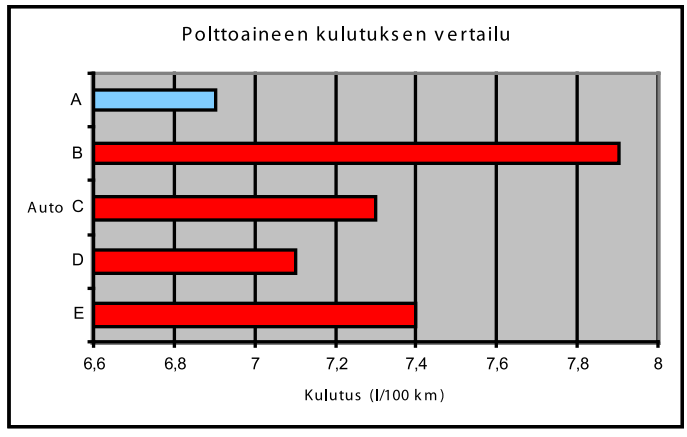
	2000	2001	2002	2003	2004	2005	2006	2007
Neliöhinta	1928	1886	2038	2163	2346	2412	2693	2944
Elinkustannusindeksi	1476	1522	1551	1580	1581	1585	1607	1642

Mieti, miksi asuntojen hinnat ovat pääkaupunkiseudulla kasvaneet muita elinkustannuksia nopeammin 2000-luvulla.

Grafiikka voi myös vääristää totuutta

ESIMERKKI 14

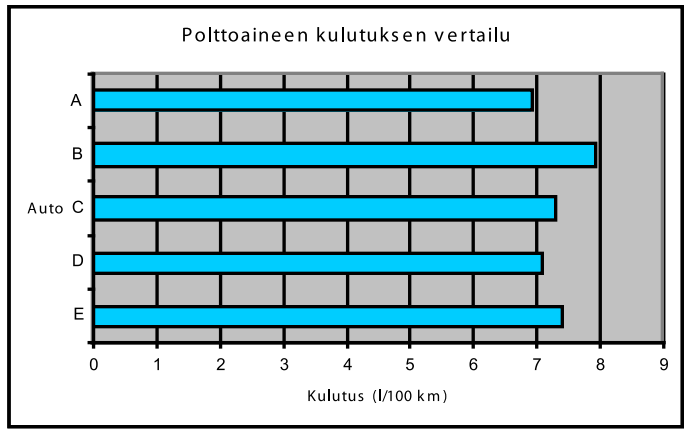
Automerkin A polttoaineen kulutusta verrattiin kilpailevien merkkien kulutukseen seuraavan kuvion avulla:



Miksi kuviosta ei saa oikeata kuvaa autojen polttoaineen kulutusten eroista?

Esimerkiksi auton B polttoaineen kulutus näyttää moninkertaiselta autoon A verrattuna, vaikka todellisuudessa ero on vain yksi litra 100 kilometriä kohden. Pääsyy virheelliseen mielikuvaan on, että asteikko on katkaistu siten, että se alkaa 6,6 litrasta. Eroa korostaa palkkien erilainen väritys. Kulutusten visuaalinen vertailtavuus ei ole kunnossa. Lisäksi auton A vierelle on sijoitettu vertailuautoista kaikkein eniten kuluttava. Mainonnassa käytetään usein tämän kaltaisia temppuja.

Seuraavassa sama kuvio on piirretty asianmukaisesti. Vertailu paljastaa eron totuuden ja vääristelyn välillä: autojen polttoaineen kulutuksen erot ovat todellisuudessa vähäiset.



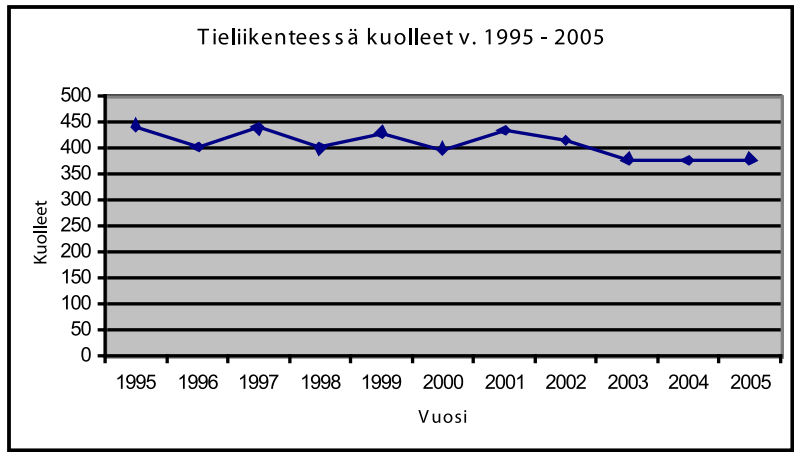
ESIMERKKI 15

Suomessa tieliikenteessä kuolleiden määrä vuosina 1995–2006 oli seuraava (lähde: Tilastokeskus):

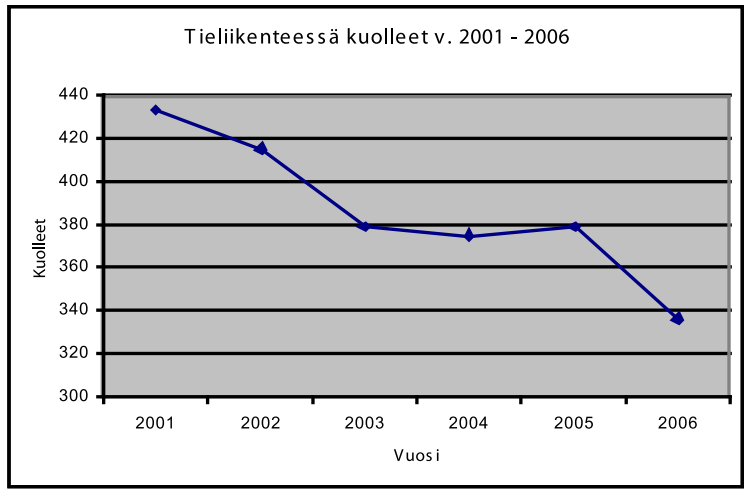
Vuosi	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Kuolleet	441	404	438	400	431	396	433	415	379	375	379	336

Seuraavat kaksi viivadiagrammia kuvaavat molemmat edellä esitettyä aikasarjaa. Kuvaajista välittyy kuitenkin hyvin erilainen kuva tieliikenneonnettomuuksissa kuolleista. Ero johtuu erilaisista arvoasteikkojen (y-akselien) mittayksikkövalinnoista ja tilastoon valituista vuosista.

Kuva 1



Kuva 2



Kuvan 1 perusteella näyttää, että liikennekuolemien määrä on pysynyt vuosittain lähes ennallaan. Kuvasta 2 saa puolestaan käsityksen, että liikennekuolemien määrä on vähentynyt melko nopeasti.

Muun muassa edellä olevien esimerkkien perusteella eri yhteyksissä esitettyä grafiikkaa on aina syytä tarkastella kriittisesti. Noudattamalla tiettyjä yksinkertaisia ohjeita on mahdollista laatia kuvio, joka antaa objektiivisen kuvan tarkasteltavasta ilmiöstä.

Kuvioiden laatimisohjeita

Jotta edellä mainitut tavoitteet saavutettaisiin, kannattaa noudattaa seuraavia ohjeita:

1. Kuvion tulee kertoa totuus esitettävästä asiasta.
2. Tee kuviosta mahdollisimman yksinkertainen, mutta ota kuitenkin edellinen ohje huomioon.
3. Valitse kuhunkin tapaukseen parhaiten soveltuva kuviotyyppi.
4. Älä liitä kuvioon sellaisia tehosteita (mm. kolmiulotteisuus), jotka vievät huomion pääasiasta.
5. Käytä tarpeen mukaan otsikoita, selitteitä ja apuviivoja ja merkitse asteikkoihin mittayksikkö. Otsikko antaa yleensä vastauksen kysymyksiin mitä, missä, milloin.
6. Käytä värejä tai varjostuksia elävöittämään kuvion sanomaa.
7. Tulkitse aina kuvion sanoma ja siitä tehtävät johtopäätökset myös sanallisesti tilastoraporttiin.

TEHTÄVIÄ

1. Tilastoaineiston pienin ja suurin arvo ovat

a) 141 cm ja 198 cm

b) 455 € ja 3 210 €.

Mikä on sopiva luokkaväli ja ensimmäisen luokan alaraja, jos luokittelussa käytetään kuutta luokkaa?

2. Yrityksen työntekijöiden kokonaiset poissaolopäivät olivat eräänä vuonna seuraavat:

5	3	16	9	0	12	3
2	19	5	19	11	2	0
4	6	1	15	4	0	2
4	0	2	4	3	3	8
3	9	11	19	2	0	1
1	2	16	4	3	2	7
7	4	1	9	13	6	5
8	9	0	4	2	6	13
0	7	1	2	4	3	17
9	4	5	2	1	6	7
0	2	10	3	12	7	2
0	1	5	17	3	5	6
11	10	4				

Luokittele ja taulukoi jakauma tasavälisiin luokkiin.

3. Markkinatutkimuksessa haastateltujen henkilöiden vuositulot (1 000 €) käyvät ilmi seuraavasta taulukosta:

44	33	34	19	15	22	29
47	20	10	18	17	11	25
12	23	29	26	19	17	16
28	14	20	49	22	16	24
47	39	14	19	27	18	39
24	54	29	42	11	42	51
22	17	23	21	15	34	18
22	24	33	21	13	58	21
17	27	24	30	15	28	16
38	21	13	23	41	26	39

Luokittele ja taulukoi jakauma tasavälisiin luokkiin.

4. Laaduntarkkailussa tutkittiin sähkölamppujen palamisaikaa. Otokseen poimittujen lamppujen palamisaika (100 h) ilmenee seuraavasta taulukosta:

18,3	19,3	18,6	11,2	22,2	19,7	14,0
16,2	14,3	16,8	19,9	20,4	17,3	10,3
13,7	15,3	17,0	11,9	10,9	15,5	17,2
14,7	14,5	10,6	11,9	9,8	21,2	18,2
12,9	18,0	17,7	9,9	15,5	9,7	14,0
12,9	12,8	14,4	18,3	12,6	16,8	19,0
10,6	16,7	16,9	14,8	20,1	16,5	16,0
16,6	10,9	19,7	20,0	13,0	19,0	10,9
16,2	15,5	17,0	16,7	9,9	21,5	20,0
20,3	13,2	14,5	13,3	20,0	17,7	9,3
17,5	15,5	13,9	16,3	19,6	14,7	19,2
16,7	11,0	13,9				

Luokittele ja taulukoi jakauma tasavälisiin luokkiin.

5. Luokittele ja taulukoi havaintomatriisin ASIAKAS muuttuja 2 (ikä). Merkitse taulukkoon myös suhteelliset frekvenssit ja suhteelliset kumulatiiviset frekvenssit.

6. Luokittele ja taulukoi havaintomatriisin ASIAKAS muuttujat
1 (sukupuoli)
3 (kotitalouden koko)
4 (asuminen)
7 (kotitalouden päivittäistavaroiden ostot)
9 (tyytyväisyys liikkeen hintatasoon).
Määritä myös kumulatiiviset frekvenssit silloin, kun se on järkevää.

7. Tutki havaintomatriisin ASIAKAS seuraavien muuttujien riippuvuutta ristiintaulukoinnin avulla.
a) 3 (kotitalouden koko) ja 9 (tyytyväisyys liikkeen hintatasoon)
b) 1 (sukupuoli) ja 8 (tyytyväisyys liikkeen tuotevalikoimaan)

8. Asiakastutkimuksessa, johon havaintomatriisi ASIAKAS liittyy, haluttiin selvittää, onko tyytyväisyydellä liikkeen tuotevalikoimaan ja tyytyväisyydellä hintatasoon riippuvuutta? Tee johtopäätökset ristiintaulukoimalla muuttujat.

9. Tutkimuksessa haluttiin selvittää eri-ikäisten ja eri sukupuolta olevien henkilöiden suhtautumista varusmiespalvelukseen. Tutkittavat muuttujat olivat seuraavat:

1 sukupuoli (1 = nainen, 2 = mies)
2 ikä (vuosina)
3 suhtautuminen varusmiespalvelukseen (1–5; 1 = erittäin kielteinen, 5 = erittäin myönteinen).

Havaintoarvot tilastoyksiköittäin on esitetty seuraavassa taulukossa:

1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	26	3	2	37	3	1	50	1	2	59	3	2	31	4
1	36	3	2	41	4	1	56	4	1	26	2	1	32	2
2	36	4	2	44	5	2	44	4	1	45	3	2	44	3
2	58	5	2	49	4	2	40	5	2	30	3	1	37	3
2	28	4	2	60	4	2	22	3	1	28	4	2	24	5
1	62	3	2	50	5	2	28	4	2	45	4	1	48	3
1	21	1	2	40	5	2	26	1	1	32	3	1	26	2
2	55	5	2	22	2	2	33	5	2	42	4	2	30	4
2	62	4	2	34	4	2	21	1	2	26	3	1	37	2
2	29	4	1	32	2	1	60	3	2	49	5	2	30	2
1	59	4	1	37	3	1	53	3	1	47	3	1	28	2
2	30	1	1	38	3	1	27	3	1	46	1	2	45	4
1	41	3	2	58	3	2	60	3	1	29	3	2	39	3
1	30	1	1	38	4	2	33	3	2	20	1	2	20	1
2	27	4	1	41	3	1	45	2	1	27	2	2	61	4
1	41	4	2	58	5	2	60	5	1	54	3	1	40	2
2	38	5	2	56	4	2	27	3	2	62	5	2	38	3
2	40	5	2	60	4	2	60	4	1	38	3	1	30	2
1	40	1	1	24	3	2	35	3	2	40	4	1	49	4
2	22	3	2	55	4	1	28	2	2	49	5	2	32	3
1	20	1	1	34	3	2	56	5	2	31	1	2	48	4
1	20	2	2	31	4	2	35	2	1	25	1	1	33	2
1	45	2	1	29	1	2	31	4	1	20	2	2	21	2
1	50	2	2	60	3	1	46	2	2	38	4	1	41	3
1	59	1	1	29	2	2	26	2	2	28	1	1	23	3

Tutkimuksen tarkoituksena oli selvittää

a) taustamuuttujien 1 (sukupuoli) ja 2 (ikä) jakaumat
b) riippuuko suhtautuminen varusmiespalvelukseen sukupuolesta
c) onko iällä ja suhtautumisella varusmiespalvelukseen riippuvuutta.

Taulukoi muuttujat ja pyri tekemään sen jälkeen johtopäätökset.

10. Bruttokansantuotteen muutosprosentti edelliseen vuoteen verrattuna vuosina 2000–2006 oli seuraava (lähde: Tilastokeskus):

Vuosi	Bruttokansantuotteen muutos (%)
2000	5,0 %
2001	2,6 %
2002	1,6 %
2003	1,8 %
2004	3,7 %
2005	2,9 %
2006	5,0 %

Esitä tilasto
a) pylväsdiagrammina,
b) viivadiagrammina.

Millaisia johtopäätöksiä siitä voi tehdä?

11. Suomalaisten internetin ja tietokoneen käyttöön viikossa kuluttama aika (15–74-vuotiaat) vuonna 2007 oli seuraava (lähde: Tilastokeskus):

Aika (h/viikko)	Naiset	Miehet
Alle 1 h	17 %	12 %
$1 \leq$ aika < 2	16 %	13 %
$2 \leq$ aika < 3	14 %	8 %
$3 \leq$ aika < 7	27 %	28 %
$7 \leq$ aika < 14	17 %	22 %
$14 \leq$ aika < 21	6 %	11 %
Vähintään 21	3 %	6 %

a) Esitä pylväsdiagrammina erikseen naisten ja miesten internetin ja tietokoneen käyttöön kuluttama aika.
b) Esitä naisten ja miesten internetin ja tietokoneen käyttöön kuluttama aika samassa kuviossa.
c) Mitä tilastosta voidaan päätellä?

12. TNS-Gallup tutki vuonna 2007, kuinka hyvin asiakkaat saavat yrityksiin yhteyden internetin kautta. Kokonaan vastaamattomien viestien määrä toimialoittain oli seuraava (lähde: TNS-Gallup):

Toimiala	Vastaamattomien viestien osuus (%)
Teleliikenne	33
Vähittäismyynti	15
Matkamyynti	15
Lehtimyynti	39
Vakuutusyhtiöt	5
Pankit	15
Liikenne	24
Hotellivaraamot	26
Automyynti	32
Energiayhtiöt	36
Elektroniikan vähittäismyynti	24
Lippupalvelut	90
Huonekalumyynti	49
Pelipalvelut	18
Autovuokraamot	17
Asunnonvälitys	40
Laajakaistapalvelut	30
Kanta-asiakaspalvelut	8

Esitä tilasto sopivana pylväskuviona ja tulkitse kuvion sanoma.

13. Korkean teknologian tuotteiden osuus Suomen koko tuonnista ja viennistä vuosina 1997–2006 oli seuraava (lähde: Tullihallitus):

Vuosi	Osuus tuonnista (%)	Osuus viennistä (%)
1997	15,4	16,1
1998	16,9	19,0
1999	18,0	20,5
2000	18,9	23,3
2001	17,7	20,9
2002	16,0	20,6
2003	14,7	20,1
2004	14,2	17,5
2005	15,6	21,3
2006	14,1	18,2

Esitä tuonnin ja viennin osuudet samassa kuviossa
a) pylväsdiagrammina,
b) viivadiagrammina.

14. Esitä graafisesti havaintomatriisiin ASIAKAS muuttujien 8 (tyytyväisyys liikkeen tuotevalikoimaan) ja 9 (tyytyväisyys liikkeen hintatasoon) jakaumat.

15. Esitä havaintomatriisin ASIAKAS muuttujan 3 (kotitalouden koko) jakauma graafisesti
a) pylväsdiagrammina,
b) sektoridiagrammina.

16. Nordean Osakesalkku -sijoitusrahaston osakkeiden toimialajakauma 31.7.2007 on esitetty alla olevassa taulukossa. Esitä osakkeiden toimialajakauma sektoridiagrammina.

Toimiala	Osuus salkusta (%)
Rahoitus	19,46
Teollisuustuottet ja palvelut	15,21
It-ala	11,96
Kulutustavarat ja palvelut	10,67
Terveydenhuolto	8,78
Energia	7,82
Perusteollisuus	6,53
Päivittäistavarat	5,62
Muut	13,95

17. Suomen rahalaitosten euromääräiset lainat euroalueen kotitalouksille käyttötarkoituksen mukaan olivat 30.6.2006–31.5.2007 seuraavat (lähde: Suomen Pankki):

Ajankohta	Asuntolainat (mrd. €)	Kulutusluotot (mrd. €)	Muut luotot (mrd. €)
30.6.2006	51,73	9,94	11,54
31.7.2006	52,22	10,05	11,61
31.8.2006	52,88	10,14	11,72
30.9.2006	53,67	10,23	11,87
31.10.2006	54,29	10,32	11,86
30.11.2006	54,85	10,37	12,05
31.12.2006	55,31	10,42	12,23
31.1.2007	55,69	10,44	12,27
28.2.2007	56,19	10,49	12,33
31.3.2007	56,78	10,63	12,46
30.4.2007	57,29	10,62	12,51
31.5.2007	58,04	10,74	12,61

Esitä tilasto

a) ositettuina pylväinä,

b) aluekuviona siten, että luottojen yhteismäärä on kokonaisuus, joihin luottolajeja verrataan.

18. Suomalaisten matkapuhelinten suurimmat vientimaat vuonna 2006 olivat seuraavat (lähde: Tullihallitus):

Valtio	Vienti (milj. kpl)
Iso-Britannia	3,4
Saksa	3,3
Venäjä	3,2
Kiina	2,6
Italia	2,5
Arabiemiirikunnat	2,1
Saudi-Arabia	2,1
Espanja	1,5
Turkki	1,1
Ranska	0,9

Esitä tilasto pylväsdiagrammina vaakapylväillä.

19. Työssä käyvien viikkotyöajan pituuksien suhteellinen jakauma eri Pohjoismaissa vuonna 2005 oli seuraava (lähde: Kansainvälinen työjärjestö ILO):

Valtio	Vähintään 40 h	30–39 h	20–29 h	Alle 20 h
Suomi	42 %	47 %	5 %	6 %
Ruotsi	58 %	28 %	9 %	5 %
Norja	16 %	63 %	9 %	12 %
Tanska	29 %	53 %	7 %	11 %

Esitä eri maiden viikkotyöajan suhteelliset jakaumat palkkikuviona.

20. Työn tuottavuus verrattuna Yhdysvaltain tuottavuustasoon oli ILO:n mukaan eräissä maissa vuosina 1980 ja 2005 seuraava (Yhdysvaltain tuottavuutta molempina vuosina merkittiin luvulla 100):

Valtio	Työn tuottavuus 1980	Työn tuottavuus 2005
Suomi	63	77
Italia	84	74
Hollanti	84	73
Sveitsi	87	65
Puola	27	37
Turkki	27	38
Etelä-Korea	27	59
Kiina	6	16

Esitä tilasto graafisesti ja tulkitse sen sanoma.

21. Uusiutuvien energialähteiden osuus sähkönkulutuksesta eräissä EU-maissa vuonna 2006 sekä tavoite vuoteen 2020 mennessä olivat seuraavat (lähde: Eurostat):

Valtio	Osuus v. 2006	Tavoite v. 2020
Itävalta	59 %	78 %
Ruotsi	46 %	60 %
Suomi	28 %	32 %
Tanska	27 %	29 %
Ranska	13 %	21 %
Saksa	10 %	12 %
Iso-Britannia	4 %	10 %
Puola	2 %	8 %

Esitä tilasto graafisesti ja tulkitse sen sanoma.

4
Suoran jakauman tunnusluvut

Tilastollisen muuttujan arvojen taulukointi ja graafinen esitys antavat yleiskuvan muuttujan **jakaumasta**. Kuvion tulkinta on lisäksi ainakin jossakin määrin subjektiivinen. Johtopäätösten ja objektiivisen tulkinnan tueksi tarvitaan **tunnuslukuja**, jotka lasketaan jakaumasta. Tunnuslukuja on paljon, ja jokainen niistä valottaa muuttujan jakaumaa eri näkökulmista. Kun tunnuslukujen sanomat yhdistetään, jakaumasta saadaan oleellista tietoa.

Tilasto- ja taulukkolaskentaohjelmien avulla voidaan vaivattomasti määrittää lähes mikä tahansa tunnusluku. Vaikka ohjelmat laskevat mekaanisesti halutun tunnusluvun, on tiedettävä tarkkaan, soveltuuko tunnusluku kyseiseen tapaukseen. Myös tunnuslukujen oikea tulkinta on keskeistä, koska ilman tulkintaa tunnusluku on vain numero tai jono numeroita.

Suorien jakaumien tunnusluvut voidaan jakaa **sijaintilukuihin** (measures of location), **hajontalukuihin** ja **muihin tunnuslukuihin.** Sijaintiluvut voidaan edelleen jakaa **keskilukuihin** ja **muihin sijaintilukuihin**.

Yksittäisen tunnusluvun perusteella tehdään joskus virheellisesti rohkeita päätelmiä koko jakaumasta. On tärkeää muistaa, että yksittäinen tunnusluku kuvaa jakaumaa vain yhdeltä kannalta. Vain tarkastelemalla yhtä aikaa useita tunnuslukuja voidaan jakaumasta tehdä luotettavia johtopäätöksiä.

4.1 Sijaintiluvut

Keskiluvut

Keskiluvut (measures of central tendency) osoittavat jakauman keskikohdan. Se voi olla myös kohta, jossa suuri osa havaintoarvoja sijaitsee.

Tunnusluvun valinta liittyy kiinteästi siihen, millä asteikolla tilastollinen muuttuja mitataan. Seuraavassa tarkastellaan sijaintiluvun määrittämistä eri asteikoilla.

Luokitteluasteikon muuttujan havainnoilla ei ole keskinäistä järjestystä. Havainnoille ei siksi voi laskea keskikohtaa missään järkevässä mielessä.

Moodi

Ainoa ”keskiluku”, jota voidaan käyttää luokitteluasteikon muuttujalle, on **moodi** (mode) eli tyyppiarvo (Mo). Moodi on havaintoaineiston yleisin muuttujan arvo tai se on luokka, jolla on suurin frekvenssi.

ESIMERKKI 1

Mielipidetiedustelussa erääseen väittämään saatujen vastausten jakauma oli seuraava:

Vastaus	Lukumäärä
Täysin samaa mieltä	37
Jokseenkin samaa mieltä	114
En osaa sanoa	52
Jokseenkin eri mieltä	90
Täysin eri mieltä	14

Tässä tapauksessa yleisin mielipide eli moodi = Jokseenkin samaa mieltä.

Moodin käyttö rajoittuu käytännössä luokittelu- tai järjestysasteikkoon. Moodi on epävarma sijaintiluku, varsinkin jos tilastoaineisto on pieni. Lisäksi jakaumalla voi olla useita moodeja tai sillä ei ole moodia ollenkaan.

Vaikka moodi voidaan määrittää myös järjestysasteikon muuttujalle, suositeltavinta on käyttää **mediaania**.

Mediaani

Mediaani (median, Md) on **järjestetyn** havaintoaineiston keskimmäinen havainto, jos havaintoarvoja on pariton määrä. Jos havaintoarvoja on parillinen määrä, mediaanina esitetään jompikumpi keskimmäisistä arvoista.

Mediaani siis jakaa järjestetyn havaintoaineiston kahteen osaan siten, että puolet arvoista on mediaania pienempiä ja puolet arvoista mediaania suurempia.

ESIMERKKI 2

a) Järjestyslukujen 2, 3, 3, 4, 8, 8, 12, 23, 25 mediaani on viides havaintoarvo eli 8.
b) Järjestyslukujen 8, 12, 14, 17, 17, 24 mediaani on joko kolmas tai neljäs havainto eli 14 tai 17.

Jos parillisen havaintojen määrän tapauksessa kaksi keskimmäistä havaintoa ovat hyvin kaukana toisistaan, on aiheellista ilmoittaa mediaaniksi kumpikin keskimmäinen havainto tai jättää mediaani kokonaan ilmoittamatta.

ESIMERKKI 3

Haastattelussa pyydettiin ihmisiä ottamaan kantaa väitteeseen ”Suomeen ei tulevaisuudessa pitäisi rakentaa lisää ydinvoimaa” (1 = täysin eri mieltä, 2 = jokseenkin eri mieltä, 3 = ei mielipidettä tai neutraali kanta, 4 = jokseenkin samaa mieltä, 5 = täysin samaa mieltä).

Haastattelun avulla saatu havaintoaineisto oli seuraava:

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5.

Tässä aineistossa mediaani olisi ”1” tai ”5” tai ”1 ja 5”. Kaikissa tapauksissa mediaani on huono keskiluku kuvaamaan näin **polarisoitunutta** eli **kaksinapaista** aineistoa.

Mediaani on erityisesti järjestysasteikolla mitatulle muuttujalle sopiva keskiluku, mutta sillä on käyttöä myös välimatka- tai suhdeasteikolla mitatulle muuttujalle, varsinkin jos jakauma ei ole symmetrinen.

Välimatka- tai suhdeasteikolla mitatun muuttujan mediaani on keskimmäisten arvojen aritmeettinen keskiarvo, jos havaintoarvoja on parillinen määrä.

Esimerkissä 2b) mediaaniksi voidaan ilmoittaa 15,5, jos muuttuja on mitattu välimatka- tai suhdeasteikolla.

Luokitellun tilastoaineiston mediaani voidaan laskea välimatka- ja suhdeasteikolla seuraavalla laskukaavalla:

Luokitellun tilastoaineiston mediaani

$$Md = L_{Md} + \frac{\frac{n}{2} - F_{Md-1}}{f_{Md}} \cdot c_{Md}\text{, jossa}$$

L_{Md} = mediaaniluokan todellinen alaraja

F_{Md-1} = mediaaniluokkaa edeltävän luokan kumulatiivinen frekvenssi

f_{Md} = mediaaniluokan frekvenssi

c_{Md} = mediaaniluokan luokkaväli

n = havaintoarvojen lukumäärä.

ESIMERKKI 4

Tutkimuksessa selvitettiin kotitalouksien asumiskustannuksia (euroa/vuosi). Asumiskustannusten jakauma oli seuraava:

Asumiskustannukset (€/vuosi)	*f*	*F*
800–1 190	23	23
1 200–1 590	41	64
1 600–1 990	78	142
2 000–2 390	33	175
2 400–2 790	45	220
2 800–5 490	18	238
Yhteensä	**238**	**–**

Laske vuotuisten asumiskustannusten mediaani.

Koska $n = 238$, niin $\frac{n}{2} = 119$. Mediaani kuuluu siis luokkaan, jonka kumulatiivinen frekvenssi on vähintään 119. Mediaanin sisältävä luokka on 1 600–1 990 €.

$L_{Md} = 1\,595$ €; $F_{Md-1} = 64$; $f_{Md} = 78$ ja $c_{Md} = 400$ €.

$$Md = 1\,595\text{ €} + \frac{119 - 64}{78} \cdot 400\text{ €} = 1\,877\text{ €}$$

Tulos tarkoittaa sitä, että 50 %:lla kotitalouksista vuotuiset asumiskustannukset olivat pienemmät ja 50 %:lla suuremmat kuin 1 877 €.

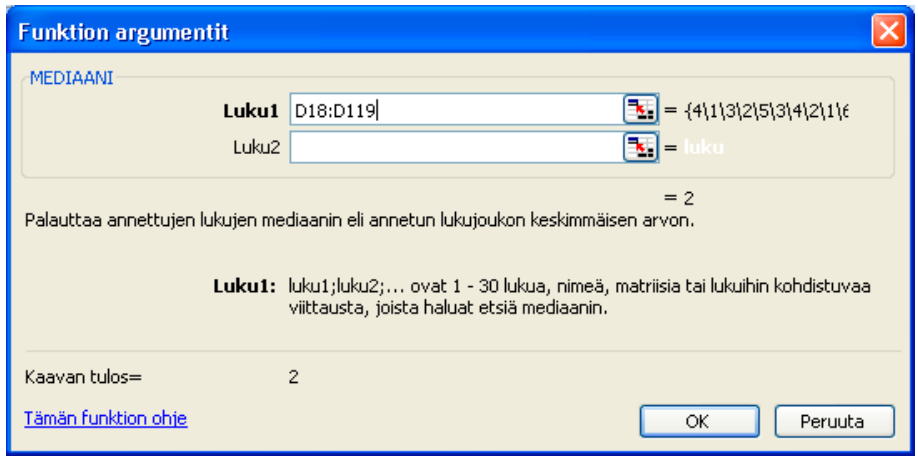
Järjestysasteikon lisäksi mediaania tulisi käyttää myös välimatka- ja suhdeasteikoilla, kun jakauma on hyvin epäsymmetrinen (esim. monet palkkatilastot) tai jos yksi tai muutama havaintoarvo poikkeaa muista arvoista huomattavasti. Mediaani voidaan laskea, vaikkei luokittelu ole tasavälinen, kuten edellinen esimerkki osoitti. Mediaani voidaan laskea myös, kun luokittelu sisältää avoimia luokkia.

Käytännössä mediaani määritetään lähes aina alkuperäisestä havaintoaineistosta tilasto-ohjelman avulla. Myös Excelissä on valmis funktio MEDIAANI (engl. MEDIAN), jonka avulla mediaani voidaan laskea. Funktio laskee mediaanin luokittelemattomasta aineistosta.

ESIMERKKI 5

Määritä havaintomatriisista ASIAKAS muuttujan ”Kotitalouden (perheen) koko” mediaani Excelin avulla.

Perheen koko on havaintomatriisin muuttuja 3. Muuttujan arvot ovat soluissa D18…D119.



Kaavan tulos on 2, eli perheen mediaanikoko on kaksi henkilöä.

Eniten käytetty välimatka- tai suhdeasteikon keskiluku on kuitenkin **aritmeettinen keskiarvo** (mean).

Aritmeettinen keskiarvo

Havaintoarvojen $x_1, x_2, \ldots x_n$ aritmeettinen keskiarvo (arithmetic mean tai mean) $\overline{x}$ määritellään seuraavasti:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}, \text{ jossa}$$

n = havaintojen lukumäärä.

Jos havaintoarvo x_1 esiintyy f_1 kertaa, havaintoarvo x_2 esiintyy f_2 kertaa,… ja havaintoarvo x_k esiintyy f_k kertaa, on keskiarvo

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}, \text{ jossa}$$

$n = \sum_{i=1}^{k} f_i$ = havaintoarvojen lukumäärä. (arithmetic mean tai mean)

Huomautus

$\sum_{i=1}^{n} x_i$ luetaan: summa x_i, jossa indeksi i saa arvot yhdestä n:ään.

Se tarkoittaa summaa $x_1 + x_2 + ... + x_n$.

ESIMERKKI 6

Laske esimerkin 4 asumiskustannusten jakauman aritmeettinen keskiarvo taulukkoa käyttämällä.

Koska tilastoaineisto on luokiteltu, jokaista luokkaa edustamaan valitaan kyseisen luokan luokkakeskus (taulukossa sarake x_i).

Asumiskustannukset (€/vuosi)	f_i	x_i	$f_i x_i$
800–1 190	23	995	22 885
1 200–1 590	41	1 395	57 195
1 600–1 990	78	1 795	140 010
2 000–2 390	33	2 195	72 435
2 400–2 790	45	2 595	116 775
2 800–5 490	18	4 145	74 610
Yhteensä	**238**		**483 910**

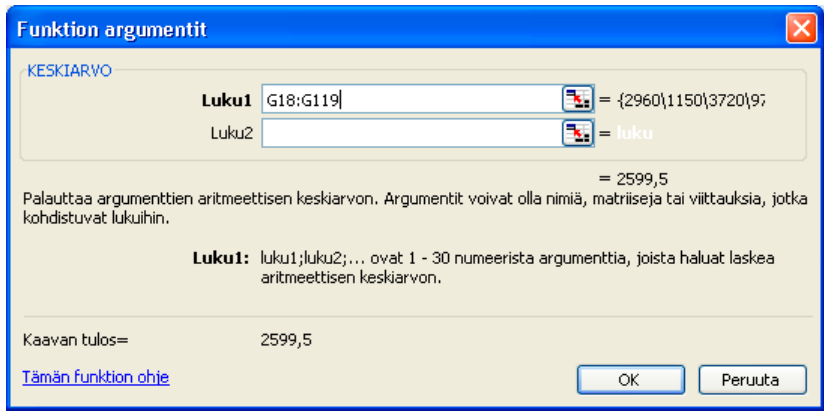
$$\sum_{i=1}^{6} f_i x_i = 483\,910;\ n = \sum_{i=1}^{6} f_i = 238;\ \bar{x} = \frac{483\,910}{238}\ € / kk = 2\,033\ €/\text{kk}$$

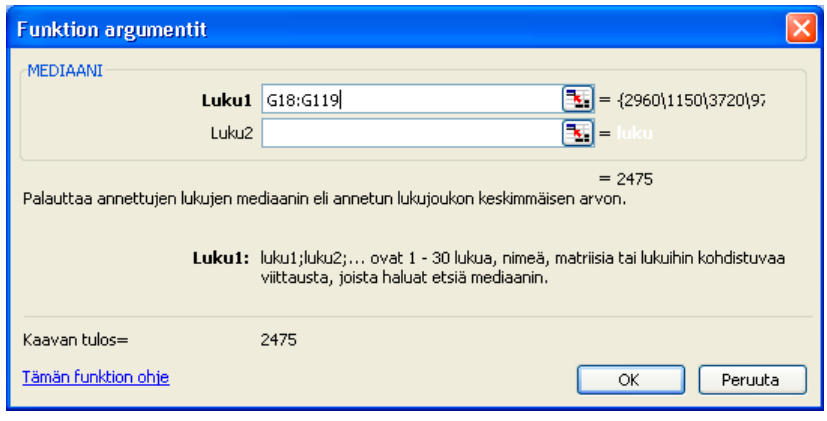
Vaikka keskiarvon ja mediaanin avulla mitataan periaatteessa samaa asiaa, tässä tapauksessa jakauman keskiarvo ja mediaani eroavat toisistaan melkoisesti. Asumiskustannusten keskiarvo on yli 150 euroa mediaania suurempi. Ero johtuu siitä, että keskiarvoon vaikuttavat muista selvästi poikkeavat havaintoarvot. Viimeisen luokan yläraja 5 490 euroa osoittaa, että muutamien kotitalouksien asumiskustannukset ovat huomattavan suuret. Keskiarvon käyttöä sijaintilukuna tulee tällöin välttää. Keskiarvoa ei voida lainkaan laskea, kun luokittelu sisältää **avoimia luokkia**.

Luokitellusta aineistosta laskettu keskiarvo on aina likiarvo, koska ei tiedetä tarkkaan, kuinka havaintoarvot kunkin luokan sisällä ovat jakautuneet. Tilasto- ja taulukkolaskentaohjelmien laskettaessa käytetään luokittelematonta aineistoa.

ESIMERKKI 7

Laske havaintomatriisin ASIAKAS muuttujan ”Kotitalouden nettotulot” keskiarvo ja mediaani Excelin avulla.





ASIAKAS		
Muuttuja 3		
	Keskiarvo	2 600
	Mediaani	2 475

Tässäkin tapauksessa keskiarvo ja mediaani eroavat toisistaan nettotulojakauman vinouden vuoksi.

Kumpi tunnusluku, keskiarvo vai mediaani, on parempi keskiluku kuvaamaan havaintosarjojen keskimääräistä arvoa? Yksikäsitteistä vastausta ei voida antaa, vaan voidaan vain muistaa molempien keskilukujen ominaiset piirteet:

- Kaikki havainnot vaikuttavat keskiarvoon, siis myös yksittäiset ”todella suuret” tai ”todella pienet” havainnot.
- Mediaani on täysin immuuni ääriarvoille, sillä sen arvoon vaikuttavat vain havaintoaineiston keskimmäinen arvo tai keskimmäiset arvot.
- Keskiarvoa pitäisi välttää tunnuslukuna, jos muutama arvo poikkeaa huomattavasti valtaosasta muita arvoja tai aineisto on vino.

Muita sijaintilukuja

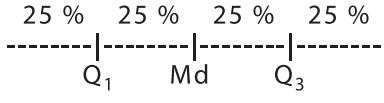
Luokitteluasteikolla mitattaville muuttujille ei ole olemassa luontevia sijaintilukuja keskilukujen yhteydessä esitetyn moodin lisäksi.

Järjestysasteikolla mitatuille muuttujille voidaan mediaanin lisäksi määrittää **fraktiileja** (fractile). Ne ovat sijaintilukuja, joita käytetään yhdessä mediaanin kanssa antamaan lisätietoa jakaumasta.

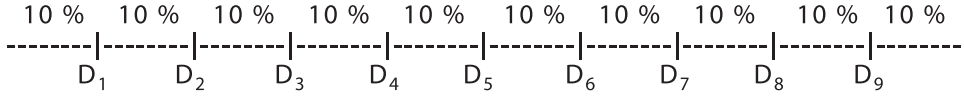
Fraktiilit

P %:n fraktiili jakaa tilastoaineiston kahteen osaan siten, että kyseistä fraktiilia pienempiä havaintoarvoja on p %.

Eniten käytettyjä fraktiileja ovat **kvartiilit** (quartile). **Alakvarttiili** Q_1 on 25 %:n fraktiili, ja **yläkvartiili** Q_3 on 75 %:n fraktiili.



Tietyistä fraktiileista käytetään nimitystä **desiili** (decile). Ensimmäinen desiili D_1 on 10 %:n fraktiili,… ja yhdeksäs desiili D_9 on 90 %:n fraktiili.



ESIMERKKI 8

Tavaratalon myyntihenkilökunnan englanninkielen taitoa haluttiin kehittää. Aluksi testattiin myyntihenkilökunnan silloinen englanninkielen taito, minkä jälkeen henkilökunta jaettiin seitsemään kielitaitoryhmään (0 = ei osaa englantia lainkaan, 1 = hallitsee yksinkertaista englanninkieltä auttavasti,... ,5 = käyttää ja ymmärtää englantia monipuolisesti ja virheettömästi, 6 = äidinkieltä vastaava englanninkielen hallinta).

Tavaratalon 120 myyjän luokitukset ilmenevät seuraavasta taulukosta:

Taso	*f*	*F*
0	6	6
1	13	19
2	28	47
3	33	80
4	22	102
5	16	118
6	2	120

Ilmoita aineistosta a) Q_1, b) D_9.

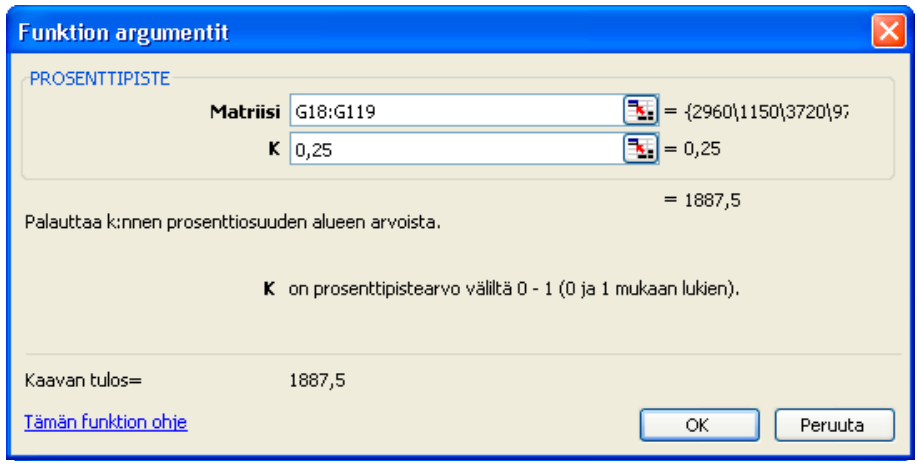
a) 25 % 120:stä on 30. Q_1 on aineiston 30. havainto. Kumulatiivisen frekvenssin perusteella havaitaan, että $Q_1 = 2$.
b) 90 % 120:stä on 108. Kumulatiivisen frekvenssin perusteella $D_9 = 5$.

Edellä mainittujen sijaintilukujen perusteella voidaan ilmoittaa, että heikoimmin englantia hallitseva neljännes sijoittuu tasoille 0–2. Paras kymmenesosa sijoittuu tasoille 5–6.

Fraktiilit voidaan määrittää myös välimatka- tai suhdeasteikolla mitatulle muuttujalle. Manuaalisesti laskettaessa sovelletaan samantapaista kaavaa kuin mediaania laskettaessa. Käytännössä laskut tehdään tilasto- tai taulukkolaskentaohjelmilla. Minkä tahansa prosenttipisteen voi määrittää myös Excelin funktion PROSENTTIPISTE(matriisi; k) (engl. PERCENTILE) avulla. Matriisi on alue, jossa havaintoarvot sijaitsevat, ja k on prosenttiluku desimaalimuodossa. Jos haluaa laskea tilastoaineiston ensimmäisen desiilin D_1 arvon, muuttujalle *k* annetaan arvo 0,10.

ESIMERKKI 9

Määritä havaintomatriisin **ASIAKAS** muuttujan 6 (kotitalouden nettotulot) alakvartiili Excelin funktion avulla.



Kotitalouden nettotulot ovat 25 prosentilla kotitalouksista enintään 1 887,50 €/kk. Vastaavasti 75 prosentilla kotitalouksista ne ovat vähintään 1 887,50 €/kk.

4.2 Hajontaluvut

Hajontaluvuilla mitataan, kuinka laajalle tai suppealle välille havaintoarvot sijoittuvat ja usein myös sitä, kuinka tiheästi havaintoarvot ovat sijoittuneet keskiluvun ympärille. Hajontaa voidaan kuvata seuraavilla kaavioilla:

Pieni hajonta:

Suuri hajonta:

Hajontalukuja tarvitaan paitsi kuvaamaan havaintoaineiston homogeenisuutta tai heterogeenisuutta myös tilastollisessa päätöksenteossa (ks. lukua 7).

Luokitteluasteikolla mitattavalle muuttujalle ei ole varsinaista hajontalukua. Eri luokkien frekvenssejä voidaan kyllä vertailla.

ESIMERKKI 10

Eräässä tutkimuksessa selvitettiin, poikkeavatko korkeakouluopiskelijoiden ja muun aikuisväestön mielipiteet toisistaan suhtautumisessa homoliittoihin. Väittämään "Hyväksyn samaa sukupuolta olevien parisuhteen" saatiin seuraavat tulokset:

	Korkeakouluopiskelijat	Muut
Täysin samaa mieltä	126	52
Jokseenkin samaa mieltä	109	46
En osaa sanoa	16	50
Jokseenkin eri mieltä	23	64
Täysin eri mieltä	9	61

Vastauksia tarkastelemalla havaitaan, että korkeakouluopiskelijoiden mielipiteet ovat keskittyneet kahteen ensimmäiseen luokkaan, kun taas muun aikuisväestön mielipiteet ovat jakautuneet melko tasaisesti kaikkiin luokkiin. Voidaankin sanoa, että mielipiteiden hajonta jälkimmäisessä tapauksessa on suurempi, vaikkei mitään tunnuslukua voida laskea.

Jos muuttuja on mitattu vähintään järjestysasteikolla, sille voidaan määrittää **vaihteluväli** ja **kvartiiliväli**.

Vaihteluväli ja kvartiiliväli

Vaihteluväli (range) kuvaa havaintoaineiston kokonaispeittoa. Vaihteluvälissä ilmoitetaan aineiston pienin havainto ja suurin havainto.

Vaihteluväli = (pienin havainto; suurin havainto).

Kvartiiliväli kuvaa havaintoaineiston keskimmäisen 50 %:n sijoittumista.

Kvartiiliväli = (Q_1, Q_3).

Tilasto-ohjelmat tulostavat yleensä jakauman suurimman ja pienimmän arvon sekä ylä- ja alakvartiilit. Niiden avulla voidaan välittömästi päätellä vaihteluväli ja kvartiiliväli.

Jos muuttuja on mitattu välimatka- tai suhdeasteikolla, voidaan jo käyttää varsin monenlaisia hajontalukuja.

Vaihteluvälin pituus ja kvartiilipoikkeama

Vaihteluvälin pituus R = suurin havainto – pienin havainto.

Kvartiilipoikkeama $Q = \frac{Q_3 - Q_1}{2}$.

Vaihteluväli, vaihteluvälin pituus, kvartiiliväli ja kvartiilipoikkeama sopivat käytettäviksi hajontalukuina silloin, kun sijaintilukuina käytetään mediaania ja kvartiileja.

Yleisin välimatka- tai suhdeasteikolla mitatun muuttujan hajontaluku on **keskihajonta** (standard deviation). Sitä käytetään yhdessä aritmeettisen keskiarvon kanssa.

Keskihajonta

Oletetaan, että muuttujan *X* havaintoarvot ovat **otos** jostakin perusjoukosta. Tällöin muuttujan arvojen keskihajonta *s* saadaan kaavasta

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}{n - 1}}$$, jossa

$\overline{x}$ = havaintoarvojen keskiarvo

n = havaintoarvojen lukumäärä.

Perusjoukon keskihajontaa laskettaessa kaavan nimittäjässä oleva $n - 1$ korvataan luvulla n. Perusjoukon keskihajontaa merkitään tavallisesti symbolilla σ (sigma). Jos havaintoarvojen määrä on vähintään 30, kaavojen antamien tulosten ero on käytännössä merkityksetön.

ESIMERKKI 11

Sokerinpussituskone oli reistaillut ja tehtaan johto halusi tietää, oliko tehtaalta päässyt markkinoille viallisesti täytettyjä sokeripusseja. Asian selvittämiseksi kahdesta keskusvarastosta otettiin yhteensä 16 kappaleen otos sokeripusseja. Otokseen joutuneiden pussien nettopainot (g) olivat 983, 976, 990, 997, 983, 969, 986, 995, 978, 996, 998, 986, 999, 988, 992 ja 979.

Kun pussituskone on kunnossa, sokeripussien keskimääräinen nettopaino on 990 g ja painojen keskihajonta 0,9 % pussien keskimääräisestä painosta.

Päättele, onko koneen reistailu aiheuttanut virheellisten pussien pääsyn markkinoille.

Lasketaan aluksi otokseen joutuneiden pussien keskimääräinen paino ja keskihajonta.

$$\sum_{i=1}^{16} x_i = 983 + 976 + \ldots + 979 = 15\,795$$

$$\overline{x} = \frac{15\,795}{16}\ \text{g} = 987{,}2\ \text{g}$$

$$\sum_{i=1}^{16} (x_i - \overline{x})^2 =$$

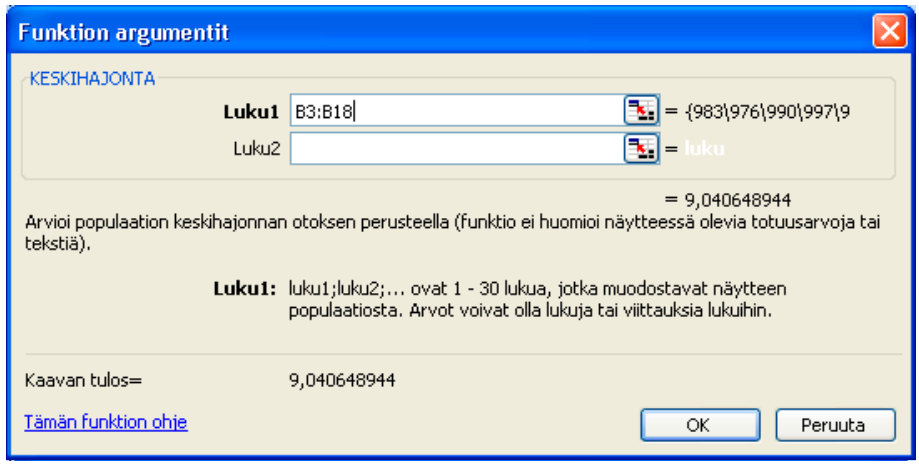
$$(983 - 987{,}2)^2 + (976 - 987{,}2)^2 + \ldots + (979 - 987{,}2)^2 = 1\,226{,}64$$

$$s = \sqrt{\frac{1\,226{,}64}{15}}\ \text{g} \approx 9{,}04\ \text{g}$$

Kun kone on kunnossa, keskihajonta on s = 0,009 · 990 g = 8,91 g. Otokseen joutuneiden pussien keskimääräinen paino oli hieman pienempi kuin 990 g. Hajonta oli hieman suurempi kuin 8,91 g. Erot ovat kuitenkin niin pieniä, ettei tämän perusteella voida päätellä, että virheellisiä pusseja olisi päässyt markkinoille.

Käytännössä muuttujan keskihajonta lasketaan aina tilasto-ohjelman tai Excelin funktion KESKIHAJONTA(luku1; luku2; …) (engl. STDEV) avulla. Luku1, Luku2… voivat olla myös solujen arvoista muodostuvia rivejä, sarakkeita tai matriiseja. Jos havaintoarvoja on vähän, keskihajonnan saa helposti lasketuksi myös useimmilla laskimilla.

Edellinen esimerkki on seuraavassa laskettu Excelin funktion avulla.



Keskihajonnan suuruutta on vaikea tulkita vertaamatta sitä joko aikaisempaan arvoon tai jonkin toisen jakauman hajontaan. Tulkintaan palataan tämän luvun lopussa.

Keskihajonnan toista potenssia s^2 sanotaan **varianssiksi** (variance). Useimmat tilasto-ohjelmat tulostavat myös sen. Matemaattisessa tilastotieteessä varianssia käytetään usein, kun johdetaan kaavoja.

Jos muuttuja on mitattu suhdeasteikolla**,** hajontalukuna voidaan käyttää **variaatiokerrointa** (coefficient of variation), joka on jakauman suhteellinen hajonta.

Variaatiokerroin

Variaatiokerroin V osoittaa jakauman suhteellisen hajonnan. Se saadaan keskihajonnan ja keskiarvon osamääränä.

$$V = \frac{s}{\overline{x}}$$

ESIMERKKI 12

Suuryrityksen myyntimiehistä valittiin 7 hengen otos. Otokseen poimittujen myynnin määrän (€/kk) ja palkkojen (€/kk) jakaumat olivat seuraavan taulukon mukaiset.

Myynti (€/kk)	40 500	48 500	33 250	58 000	42 500	60 000	51 250
Palkka (€/kk)	2 550	3000	2 160	3 550	2 725	3 825	3 190

Jakaumien keskiarvot ja hajonnat ovat seuraavat:

Myynnin määrä: $\overline{x} = 47\,714$ €/kk $\quad s = 9\,646$ €/kk

Palkka $\overline{x} = 3\,000$ €/kk $\quad s = 578$ €/kk.

Vaikka palkkojen hajonta on vain murto-osa myynnin määrän hajonnasta, variaatiokertoimet ovat lähes yhtä suuret.

Myynnin määrä: $V = \frac{9\,646}{47\,714} \approx 0{,}202 = 20{,}2\ \%$

Palkka: $V = \frac{578}{3\,000} \approx 0{,}193 = 19{,}3\ \%$

Tuloksista voidaan päätellä, että yrityksen myyntimiesten palkat on mitä ilmeisimmin jollakin tavalla sidottu myynnin määrään.

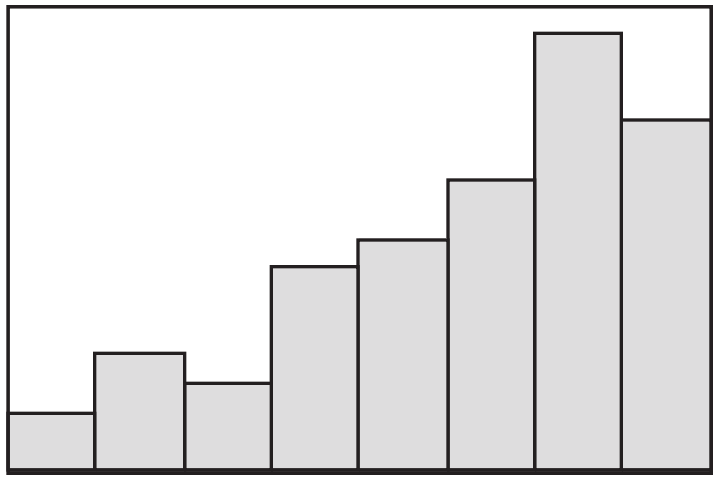
Variaatiokerrointa käytetään, kun vertaillaan kahden muuttujan jakauman hajontojen suuruutta. Vertailussa muuttujien hajontojen mittayksiköiden ei tarvitse olla samoja.

4.3 Muita tunnuslukuja

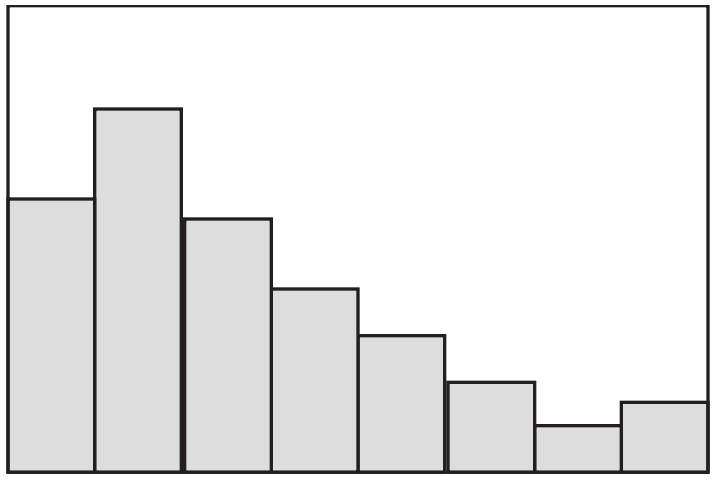
Muita muun muassa tilastollisen päätöksenteon yhteydessä käytettäviä suoran jakauman tunnuslukuja ovat **vinous** (skewness) ja **huipukkuus** (kurtosis). Vinous ja huipukkuus voidaan määrittää välimatka- tai suhdeasteikon muuttujalle. Niiden avulla arvioidaan empiirisen jakauman normaalisuutta (ks. luku 6).

Vinous ja huipukkuus

Vinous ja huipukkuus mittaavat kumpikin omalla tavallaan jakauman poikkeamaa normaalijakaumasta.



Vasemmalle (negatiivisesti) vino jakauma, jonka vinous < 0.



Oikealle (positiivisesti) vino jakauma, jonka vinous > 0.

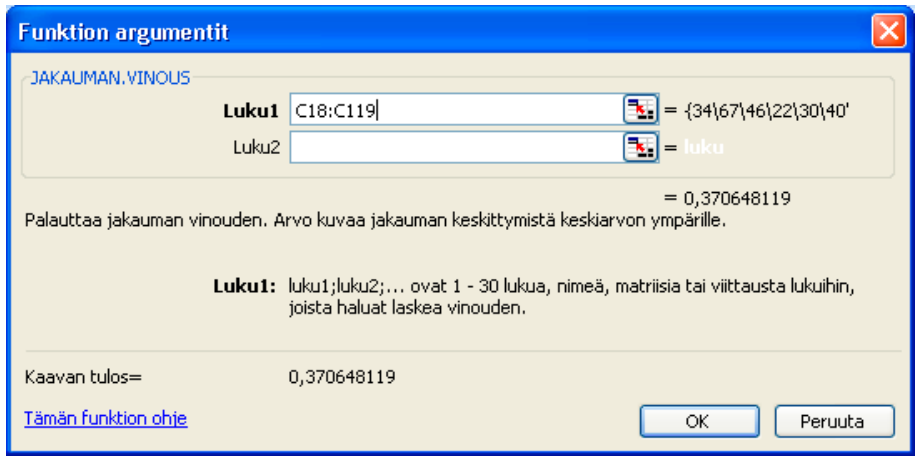
Normaalijakaumaa noudattavan muuttujan vinous on nolla. Jos vinous poikkeaa selvästi nollasta, muuttuja ei ole normaalisti jakautunut.

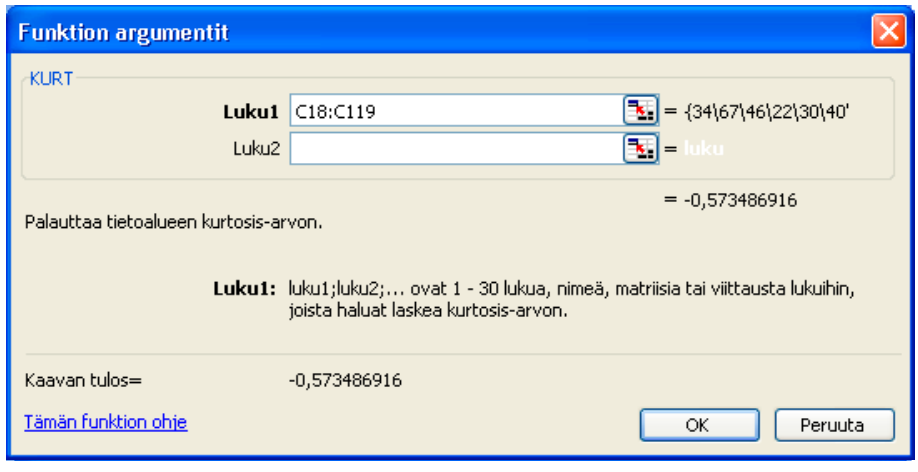
Huipukkuus mittaa jakauman huipun terävyyttä. Alkuperäistä määritelmää käyttämällä normaalijakaumaa noudattavan muuttujan huipukkuus on +3. Tilasto-ohjelmissa ja Excelissä käytetään nykyään huipukkuuden tunnuslukuna ”normitettua” huipukkuutta. Tämä tunnusluku saa arvon 0, kun muuttuja on normaalisti jakautunut. Jos huipukkuuden arvo poikkeaa selvästi nollasta, muuttuja ei ole normaalisti jakautunut.

Excelissä jakauman vinous määritetään funktion JAKAUMAN.VINOUS(Luku1;Luku2;…) (engl. SKEW) avulla ja huipukkuus funktion KURT(Luku1;Luku2;…) (engl. KURT) avulla. Argumentit Luku1; Luku2,… voivat olla lukuja tai solujen muodostamia vaakarivejä, pystyrivejä tai matriiseja.

ESIMERKKI 13

Määritä havaintomatriisin ASIAKAS muuttujan 2 (ikä) jakauman vinous ja huipukkuus.





Jakauman vinous on 0,37 ja huipukkuus –0,57.

ESIMERKKI 14

Tutkimuksessa haluttiin selvittää, poikkeavatko palvelualoilla työskentelevien naisten ja miesten palkat paikkakunnalla toisistaan. Tutkimusta varten poimittiin sadan palvelualalla toimivan työntekijän otos. Otokseen valituista naisia oli 57 ja miehiä 43. Naisten ja miesten palkkajakaumista laskettiin tilasto-ohjelman avulla seuraavat tunnusluvut:

Descriptive statistics **Naiset**	
	(€/kk)
count	57
mean	1 994,57
sample variance	135 913,14
sample standard deviation	368,66
minimum	1 593,00
maximum	3 375,00
range	1 782,00
skewness	1,54
kurtosis	2,55
coefficient of variation (CV)	0,18
1st quartile	1 741,50
median	1 890,00
3rd quartile	2 187,00
interquartile range	445,00
mode	1 741,50

Descriptive statistics **Miehet**	
	(€/kk)
count	43
mean	2 114,48
sample variance	284 214,74
sample standard deviation	533,12
minimum	1 620,00
maximum	3 766,50
range	2 146,50
skewness	1,66
kurtosis	2,18
coefficient of variation (CV)	0,25
1st quartile	1 741,50
median	1 957,50
3rd quartile	2 254,50
interquartile range	513,00
mode	1 741,50

Voidaanko tutkimuksen perusteella päätellä, että naisten ja miesten palkat poikkeavat toisistaan?

Aloitetaan tarkastelu keskiarvoista (mean). Miesten keskipalkka on tulosteen mukaan noin 120 €/kk suurempi kuin naisten keskipalkka. Tulosteesta voidaan kuitenkin havaita, että myös miesten palkkojen hajonta (sample standard deviation) ja variaatiokerroin (coefficient of variation) ovat selvästi suurempia kuin naisten palkoista lasketut vastaavat tunnusluvut. Miesten palkkojen vaihtelu on siis selvästi naisten palkkojen vaihtelua suurempaa. Naisten ja miesten pienimmät palkat (minimum) ovat lähes samat, mutta miesten suurin palkka (maximum) on noin 390 euroa kuukaudessa naisten huippupalkkaa suurempi.

Kun tarkastellaan mediaanipalkkoja (median), kvartiileja (quartile) ja kvartiilivälejä (interquartile range), havaitaan, että näiden arvojen erot ovat pienemmät kuin keskiarvojen ja keskihajontojen erot. Kun vielä todetaan, että jakaumat ovat oikealle vinoja (skewness), voidaan tehdä lopullinen tulkinta palkkaeroista.

Tulkinta: Miesten palkat ovat keskimäärin hieman naisten palkkoja suuremmat. Vähiten ansaitseva neljännes naisista ansaitsee 1 593,00–1 741,50 €/kk ja vähiten ansaitseva neljännes miehistä 1 620,00–1 741,50 €/kk. Pienimmät palkat naisilla ja miehillä ovat käytännössä samat. Keskimmäinen 50 % naisista ansaitsee 1 741,50–2 187,00 €/kk ja keskimmäinen 50 % miehistä 1 741,50–2 254,50 €/kk. Tässäkään ryhmässä erot eivät ole kovin suuret. Keskiarvopalkoissa havaittava ero syntyy suurimmaksi osaksi siitä, että ylin neljännes naisista ansaitsee 2 187,00–3 375,00 €/kk, kun vastaava neljännes miehistä ansaitsee 2 254,50–3 766,00 €/kk. Parhaiten palkatuissa tehtävissä on siis selvästi enemmän miehiä kuin naisia.

TEHTÄVIÄ

Kun teet seuraavia tehtäviä, kiinnitä erityistä huomiota laskettujen tunnuslukujen tulkintaan.

1. Tupakkatehdas tutki valmistamansa savukemerkin markkinaosuuden muutoksia. Tutkimukseen otettiin eri puolilta maata yhteensä 30 liikkeen otos. Kuukauden aikana tapahtuneen myynnin seurannan perusteella savukemerkin markkinaosuus otokseen poimituissa liikkeissä oli seuraava:

Markkinaosuus (%)					
12,5	15,7	14,1	11,9	14,8	12,9
13,6	15,4	12,0	14,2	13,9	11,3
15,0	11,8	13,3	15,2	12,7	14,3
15,2	11,9	13,7	14,2	11,7	14,4
13,0	12,9	14,8	15,0	13,3	14,0

Laske otoksesta markkinaosuuden mediaani, keskiarvo, keskihajonta ja vaihteluväli.

2. Eräässä matematiikan kokeessa arvosanojen jakauma oli seuraava:

4	5	6	7	8	9	10
1,3 %	9,8 %	15,8 %	0,3 %	23,3 %	23,4 %	6,1 %

Laske arvosanojen keskiarvo. (Yo 2001 S, lyhyt)

3. Päivittäistavaroita myyvä liike tutki, kuinka monta kertaa asiakkaat kuukauden aikana asioivat liikkeessä. Tutkimusta varten poimittiin liikkeen asiakkaista 100 henkilön otos. Otokseen valittujen henkilöiden käyntikertojen jakauma oli seuraavan taulukon mukainen:

Käyntikerrat	Henkilöitä
1–3	2
4–8	26
9–13	59
14–20	10
21–27	3

Laske käyntikertojen
a) keskiarvo ja
b) mediaani.

4. Läänin veroviranomaisten vanhojen tilastojen perusteella taksiautojen keskimääräinen ajokilometrimäärä on ollut 42 000 km vuodessa. Viranomaiset halusivat tutkia, onko taksien keskimääräinen vuosittainen ajomäärä muuttunut. Asian selvittämiseksi läänin taksiautoilijoiden joukosta poimittiin 85 autoilijan otos. Heidän autoihinsa asennettiin vuodeksi ajopiirturi, jolloin autojen ajokilometrimäärien jakaumaksi saatiin seuraavat tiedot (ajomatka 1 000 km/v):

17	34	25	48	11	37	44	32	39	30
33	38	55	28	40	22	27	12	52	41
37	66	13	33	38	59	29	35	40	19
56	30	45	31	18	37	32	21	33	37
15	38	31	45	37	57	35	60	38	49
35	46	28	25	36	24	49	32	33	39
41	38	69	31	43	35	56	35	30	44
33	32	14	48	22	41	38	29	21	33
49	47	34	36	39					

Laske jakaumasta

a) keskiarvo,
b) keskihajonta ja
c) mediaani.
d) Kuinka paljon vähintään ajoi (km/v) eniten ajanut neljännes taksiautoilijoista?

5. Yrityksen työntekijöiden poissaolopäivien määrät eräänä vuonna olivat

5	3	16	9	0	12	3	2	19	5
19	11	2	0	4	6	1	15	4	0
2	4	0	2	4	0	2	4	3	3
8	3	9	11	19	2	0	1	1	2
16	4	3	2	7	7	4	1	9	13
6	5	8	9	0	4	2	6	13	0
7	1	2	4	3	17	9	4	5	2
1	6	7	0	2	10	3	12	7	2
0	1	5	17	3	5	6	11	10	4

Määritä poissaolopäivien jakauman

a) keskiarvo,
b) keskihajonta,
c) mediaani,
d) moodi,
e) kvartiilit,
f) vaihteluväli,
g) vinous,
h) huipukkuus.

Mitä johtopäätöksiä tunnuslukujen avulla voidaan tehdä työntekijöiden poissaoloista?

6. Luokittele ja taulukoi edellisen tehtävän jakauma tasavälisiin luokkiin ja laske luokitellusta aineistosta

a) keskiarvo ja

b) mediaani ja vertaa niitä alkuperäisestä jakaumasta laskettuun keskiarvoon ja mediaaniin.

7. Laakerikuulia valmistavan tuotantolinjan laaduntarkkailu hoidettiin ottamalla jokaisesta valmistuneesta erästä 20 kuulan otos ja mittaamalla otokseen poimittujen kuulien halkaisija, jolloin saatiin seuraavat havaintoarvot (mm):

11,123	11,118	11,130	11,119	11,122
11,120	11,122	11,120	11,118	11,125
11,130	11,121	11,120	11,117	11,119
11,125	11,129	11,118	11,129	11,127

Jos vaihteluvälin pituus on enemmän kuin 0,15 % otokseen joutuneiden kuulien mediaanista, koko valmistuserä hylätään. Kuinka käy tälle erälle?

8. Markkinatutkimuksessa haastateltujen henkilöiden vuosiansiot olivat (1 000 €) seuraavat:

55	29	43	18	19	28	29
51	25	10	23	21	14	23
11	29	30	33	19	18	16
29	18	25	61	28	20	30
46	49	18	20	34	23	40
30	63	36	26	14	53	64
28	18	29	26	19	43	23
24	30	35	26	16	73	26
21	34	30	25	19	35	20
34	26	10	29	39	20	49

Määritä tilasto- tai taulukkolaskentaohjelman avulla kaikki ne tunnusluvut, joiden avulla voidaan tehdä johtopäätöksiä palkkojen jakaumasta.

9. Tarkastele havaintomatriisia ASIAKAS. Määritä tilasto- tai taulukkolaskentaohjelman avulla ne tunnusluvut, jotka kullekin muuttujalle on järkevää määrittää.

10. Mikä on havaintomatriisin ASIAKAS perusteella liikkeen asiakkaina olevien kotitalouksien
- **a)** keskimääräinen koko,
- **b)** koon mediaani ja
- **c)** tyypillinen koko.

11. Mikä on havaintomatriisin ASIAKAS perusteella
- **a)** liikkeen asiakkaan tyypillinen koulutus
- **b)** liikkeen asiakkaana olevan kotitalouden tyypillinen päivittäistavaroiden ostomäärä (€/kk)?

12. Tulkitse havaintomatriisin ASIAKAS perusteella, onko asiakaskunnan tyytyväisyys liikkeen tuotevalikoimaan vähäisempää kuin tyytyväisyys liikkeen hintatasoon vai päinvastoin. Käytä kaikkia mahdollisia tunnuslukuja, jotka valaisevat asiaa.

13. Poliisi halusi tutkia autojen renkaiden kuntoa. Tutkimusta varten järjestettiin liikenneratsia, jossa tutkittiin kaikkiaan 944 renkaan kunto mittaamalla kunkin renkaan jäljellä olevan kulutuspinnan paksuus. Jakauma oli seuraava:

Kulutuspinta (mm)	1	2	3	4	5	6	7	8
Renkaita	60	72	188	276	172	120	48	8

Laske jakauman
- **a)** mediaani,
- **b)** keskiarvo,
- **c)** keskihajonta ja
- **d)** variaatiokerroin.

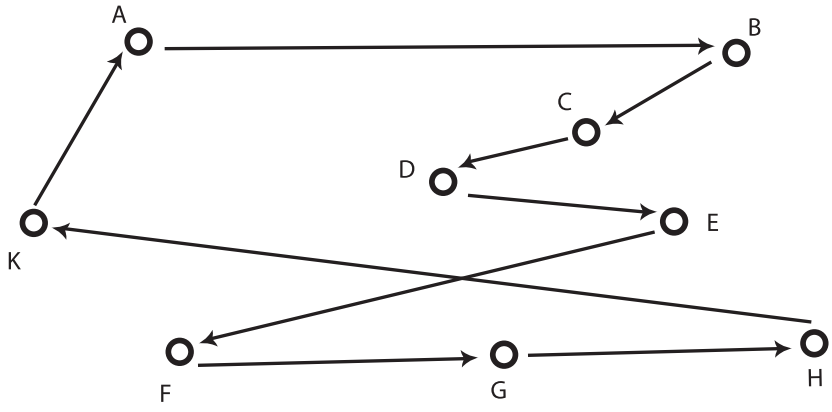
14. Mikä oli edellisessä tehtävässä kulutuspinnan paksuus enintään heikoimmalla
- **a)** neljänneksellä
- **b)** kymmenesosalla renkaista?
- **c)** Mille välille sijoittuivat kunnoltaan keskimmäiset 50 % renkaista?

15. Sähkölampunparistoja valmistava yritys tarkkaili valmistamiensa paristojen laatua ottamalla jokaisesta valmistuneesta erästä 15 pariston otoksen ja mittaamalla otokseen joutuneiden paristojen jännitteen ja jännitteiden vaihteluvälin pituuden. Edellisessä otoksessa jännitteiden keskiarvo oli 4,61 V ja vaihteluvälin pituus 0,52 V. Paristojen laatua pidetään sitä parempana, mitä pienempi vaihteluvälin pituuden suhde on keskiarvoon. Uuteen otokseen valittujen paristojen jännitteet olivat (V) 4,69; 4,82; 4,43; 4,51; 4,77; 4,66; 4,42; 4,48; 4,50; 4,56; 4,54; 4,45; 4,68; 4,49 ja 4,89.

Kumman otoksen paristojen laatu on parempi?

5
Todennäköisyyslaskenta

Tarkastellaan niin sanottua klassista kauppamatkustajan ongelmaa. Edustaja käy kuukausittain 8 paikkakunnalla. Kuinka monessa erilaisessa järjestyksessä hän voi kiertää paikkakunnat?



Kuviossa edustajan kotipaikkakuntaa on merkitty K:lla ja muita paikkakuntia kirjaimin A–H. Edustaja käy paikkakunnilla aakkosjärjestyksessä kuvioon piirrettyjen reittien mukaisesti, mikä ei selvästikään ole lyhin vaihtoehto. Jo silmämääräisesti voidaan havaita, että lyhyempi olisi esimerkiksi reitti, jossa edustaja menisi paikkakunnalta E paikkakunnalle H ja sieltä G:n ja F:n kautta kotipaikkakunnalleen.

Palataan tämän sinänsä yksinkertaisen ongelman ratkaisuun heti, kun kombinaatio-opin perusteita on käsitelty.

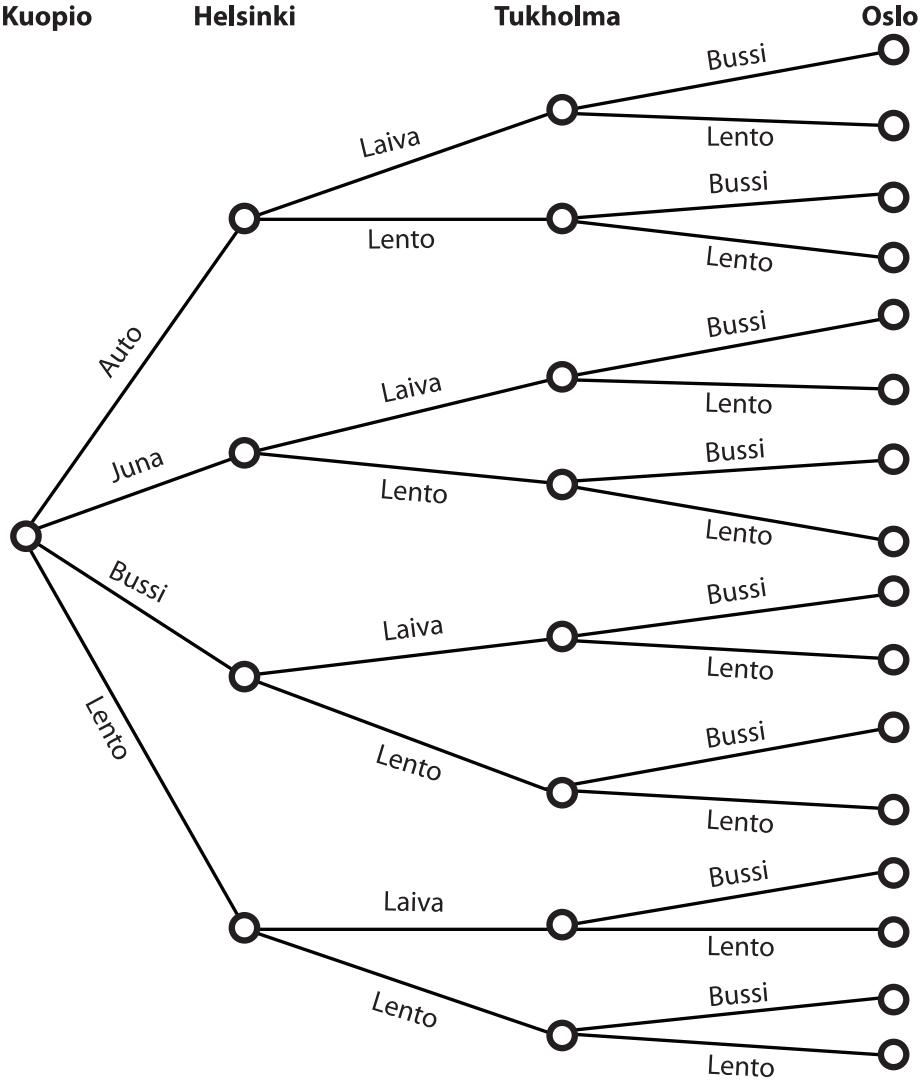
5.1 Kombinaatio-oppia

Yksi todennäköisyyslaskennan osa-alue on kombinaatio-oppi, jossa tutkitaan tietyn tapahtuman tai usean tapahtuman erilaisia toteutumismahdollisuuksia. Esimerkiksi vakuutustoiminnassa halutaan selvittää tiettyyn tapahtumaan liittyviä riskejä. Tällöin joudutaan muun muassa selvittämään, kuinka monella eri tavalla tapahtuma voi toteutua. Kombinaation käsite liittyy ennen kaikkea binomijakaumaan, jota käsitellään myöhemmin tässä luvussa.

ESIMERKKI 1

Henkilö matkustaa Kuopiosta Helsingin ja Tukholman kautta Osloon. Kuopiosta Helsinkiin hän voi mennä henkilöautolla (a), junalla (j), linja-autolla (b) tai lentokoneella (l). Helsingistä Tukholmaan hän pääsee joko laivalla (la) tai lentokoneella, ja Tukholmasta Osloon hän voi matkustaa joko linja-autolla tai lentokoneella. Kuinka monella eri tavalla hän voi tehdä matkan?

Vaihtoehtoja voidaan kuvata puurakenteella.



Kun lasketaan vaihtoehdot puurakenteen lopussa, saadaan kokonaismääräksi 16 erilaista mahdollisuutta: (a,la,b); (a,la,l); (a,l,b); (a,l,l); (j,la,b); (j,la,l); (j,l,b); (j,l,l); (b,la,b); (b,la,l); (b,l,b); (b,l,l); (l,la,b); (l,la,l); (l,l,b); (l,l,l). Tulos saadaan kertomalla eri osamatkojen suoritusvaihtoehdot keskenään: $4 \cdot 2 \cdot 2 = 16$.

Tuloperiaate

Oletetaan, että sattuu peräkkäin k tapahtumaa, joista 1. tapahtuma voi sattua n_1 eri tavalla, 2. tapahtuma n_2 eri tavalla ja k. tapahtuma n_k eri tavalla. Tällöin k tapahtuman muodostama yhdistetty tapahtuma voi sattua

$n_1 \cdot n_2 \cdot ... \cdot n_k$ eri tavalla.

ESIMERKKI 2

Vakioveikkauksessa on 13 kohdetta, joissa jokaisessa on kolme vaihtoehtoa: 1, x ja 2. Kuinka monta saraketta on täytettävä saadakseen varmasti 13 oikein?

Koska jokaisessa kohteessa on 3 eri vaihtoehtoa, vaihtoehtojen kokonaismäärä on

$3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 = 3^{13} = 1\,594\,323.$

Permutaatio

ESIMERKKI 3

Yrityksen henkilöstöpäällikön on asetettava kolme työnhakijaa paremmuusjärjestykseen. Kuinka moneen eri järjestykseen hakijat voidaan asettaa?

Jos merkitään hakijoita symbolein a, b ja c, järjestykset ovat (a,b,c); (a,c,b); (b,a,c); (b,c,a); (c,a,b); (c,b,a) eli mahdollisia järjestyksiä on yhteensä 6.

Tutkitaan kysymystä hieman yleisemmin. Oletetaan, että on asetettava järjestykseen n alkiota. Ensimmäiseksi voidaan valita mikä tahansa n alkiosta. Kun ensimmäinen alkio on valittu, toiseksi voidaan valita mikä tahansa $(n - 1)$ alkiosta. Kaksi ensimmäistä voidaan siis tuloperiaatteen mukaan valita $n \cdot (n - 1)$ eri tavalla. Jatkamalla päättelyä todetaan, että alkiot voidaan asettaa järjestykseen $n \cdot (n - 1) \cdot (n - 2) \cdot ... \cdot 2 \cdot 1$ eri tavalla.

Olkoon E joukko, jossa on n erilaista alkiota. Mitä tahansa kaikkien E:n alkioiden muodostamaa järjestettyä jonoa sanotaan E:n alkioiden **permutaatioksi**. Permutaatioiden lukumäärään liittyy käsite **luvun kertoma**.

Luvun kertoma

Olkoon n positiivinen kokonaisluku. Luvun n kertomalla tarkoitetaan lukua

$n! = n \cdot (n - 1) \cdot (n - 2) \cdot \ldots \cdot 2 \cdot 1.$

Lisäksi on sovittu, että $0! = 1$.

Huomautus
Merkintä $n!$ luetaan: n kertoma.

Permutaatioiden lukumäärä

Jos joukossa E on n erilaista alkioita, joukon alkioista muodostettujen permutaatioiden lukumäärä on $n!$, joka käytännössä tarkoittaa, kuinka moneen eri järjestykseen n alkiota voidaan asettaa.

ESIMERKKI 4

Palataan luvun alussa olleeseen kauppamatkustajan ongelmaan. Edustajan piti siis kuukausittain käydä 8 eri paikkakunnalla, ja haluttiin selvittää, kuinka monessa erilaisessa järjestyksessä hän voi kiertää kyseiset paikkakunnat.

Kysymys on selvästi permutaatioiden lukumäärästä. Ensimmäiseksi paikkakunnaksi, johon hän menee kotipaikkakunnaltaan, hän voi valita minkä tahansa 8 paikkakunnasta. Kun hän on tehnyt tämän valinnan, hän voi valita toiseksi käyntikohteekseen minkä tahansa 7 jäljellä olevasta paikkakunnasta jne.

Kaikkiaan vaihtoehtojen määrä on siis
$8! = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 40\,320.$

Vaihtoehtojen määrä on yllättävän suuri. Jos ajatellaan, että edustaja valitsisi joka kuukausi erilaisen järjestyksen 11 kuukautena vuodessa, kaikkien järjestysten kiertämiseen kuluisi $\frac{40\,320}{11}$ v $\approx 3\,665$ v.

Tarvittaisiin siis erittäin monta edustajasukupolvea ennen kuin kaikki järjestykset olisi kierretty.

Kun vaihtoehtoja on paljon, permutaatioiden lukumäärä kasvaa käsittämättömän suureksi. Jos esimerkiksi pitäisi määrittää lyhin reitti, joka kulkee kaikkien Suomen kaupunkien kautta, tehtävä on haastava jopa supertietokoneille, vaikka niiden laskentanopeus on erittäin suuri.

Usein halutaan saada selville joukon E tietyn suuruiset osajoukot ottamatta huomioon alkioiden järjestystä osajoukon sisällä.

Kombinaatio

ESIMERKKI 5

Neljän henkilön työryhmä {a,b,c,d} päättää lähettää kaksi jäsentään kansainväliseen kokoukseen. Valinta suoritetaan arpomalla. Kuinka suuri on mahdollisuus, että kokoukseen osallistujat ovat a ja b?

Muodostetaan kaikki joukon {a,b,c,d} kahden alkion osajoukot, kun alkioiden järjestyksellä ei ole merkitystä. Mahdollisia järjestyksiä on yhteensä kuusi: (a,b); (a,c); (a,d); (b,c); (b,d); (c,d).

Mahdollisuus, että kokoukseen osallistujat ovat a ja b, on siten $\frac{1}{6}$.

Joukon E osajoukkoja, joissa on r erilaista alkiota mutta joissa alkioiden järjestyksellä ei ole merkitystä, sanotaan joukon E kombinaatioiksi.

Kombinaatioiden lukumäärä

Joukon E, jossa on n erilaista alkiota, r alkiota käsittävien kombinaatioiden lukumäärä on $C_n^r = \frac{n!}{r! \cdot (n-r)!}$.

Merkintä C_n^r luetaan: ”n yli r:n.” Useimmissa laskimissa on toiminto, jolla kombinaatioiden määrän saa lasketuksi suoraan. Joskus käytetään merkinnän C_n^r sijasta merkintää $\binom{n}{r}$ tai nCr.

Kombinaatioiden lukumäärä osoittaa siis, kuinka monella eri tavalla n alkiosta voidaan valita r alkion joukkoja.

ESIMERKKI 6

30 opiskelijan ryhmästä valitaan kolme edustajaa erääseen tilaisuuteen. Kuinka monta erilaista mahdollisuutta on tehdä valinta?

$n = 30; r = 3$ $\qquad C_{30}^{3} = \frac{30!}{3! \cdot (30-3)!} = \frac{30!}{3! \cdot 27!} = 4\,060$

5.2 Todennäköisyyslaskennan peruskäsitteitä

Tarkastellaan jotakin **satunnaiskoetta** tai **-ilmiötä** ε (lue: epsilon), joita ovat muun muassa rahanheitto ja viallisten tuotteiden lukumäärä jossakin valmistusprosessissa. Niille on ominaista

- kaikkien tulosmahdollisuuksien tunnettavuus
- toistettavuus
- tietynlainen säännönmukaisuus.

ESIMERKKI 7

Rahanheitossa on kaksi tulosmahdollisuutta: kruunu ja klaava. Yksittäisen heiton tulosta ei varmasti voi sanoa, mutta jos rahaa heitetään 1 000 kertaa, kruunujen lukumäärä on ilmeisesti lähellä lukua 500.

Tarkastellaan erilaisia todennäköisyyden lähestymistapoja.

Subjektiivinen lähestymistapa

Subjektiivinen lähestymistapa perustuu omaan kokemukseen tai näkemykseen. Henkilö voi esittää arvion, että huomenna sataa 60 prosentin todennäköisyydellä, tai ennustaa, että pörssikurssit nousevat seuraavan vuoden aikana 90 prosentin todennäköisyydellä.

Subjektiivisen todennäköisyyden käytön haittapuolena on, että eri henkilöillä voi olla huomattavasti toisistaan poikkeava käsitys jonkin tapahtuman esiintymistodennäköisyydestä. Subjektiivisella todennäköisyydellä ei siten ole tilastollisessa päättelyssä merkitystä.

Klassinen lähestymistapa

Satunnaiskokeen tulosmahdollisuuksia sanotaan **alkeistapahtumiksi**. Tietyt ehdot täyttävien alkeistapahtumien joukkoa sanotaan **tapahtumaksi**. **Perusjoukko** E on kaikkien mahdollisten alkeistapahtumien muodostama joukko.

Olkoon E satunnaiskokeen tai -ilmiön kaikkien alkeistapahtumien muodostama äärellinen perusjoukko ja tapahtuma A suotuisten alkeistapahtumien muodostama E:n osajoukko. Lisäksi oletetaan, että kaikki alkeistapahtumat esiintyvät yhtä usein eli ne ovat keskenään **symmetrisiä**. $n(\mathrm{E}) = n$ = kaikkien alkeistapahtumien lukumäärä ja $n(\mathrm{A}) = k$ = suotuisten alkeistapahtumien lukumäärä. Tällöin todennäköisyyttä

$$p = \frac{k}{n}$$

sanotaan tapahtuman A **klassiseksi todennäköisyydeksi**. Tapahtuman A todennäköisyyttä (probability) merkitään P(A):lla.

ESIMERKKI 8

Olkoon tarkasteltava satunnaisilmiö noppan heitto. Tällöin perusjoukko E = {1,2,3,4,5,6}. Jos halutaan tutkia, mikä todennäköisyys on saada noppaa heittämällä parillinen silmäluku, tapahtuma A muodostuu suotuisista alkeistapahtumista 2, 4 ja 6, eli A = {2,4,6}. Koska joukon A alkioiden lukumäärä $n(\mathrm{A}) = 3$ ja $n(\mathrm{E}) = 6$, kysytty mahdollisuus eli todennäköisyys on

$$\frac{n(\mathrm{A})}{n(\mathrm{E})} = \frac{3}{6} = 0{,}5\,.$$

ESIMERKKI 9

Heitetään kerran noppaa. Mikä on todennäköisyys, että saadaan

a) silmäluku 2 tai pariton silmäluku

b) luku 1, 2, 3, 4, 5, tai 6?

a) Olkoon $A = \{2\}$ ja $B = \{1,3,5\}$. Joukkojen A ja B yhdistejoukko $A \cup B = \{1,2,3,5\}$. Perusjoukko $E = \{1,2,3,4,5,6\}$. $P(A \cup B)$ tarkoittaa todennäköisyyttä saada luku 2 tai pariton luku. Koska $n(A \cup B) = 4$ ja $n(E) = 6$, on

$$P(A \cup B) = \frac{4}{6} = \frac{2}{3}.$$

Toisaalta tapahtuman A todennäköisyys $P(A) = \frac{1}{6}$ ja tapahtuman B todennäköisyys $P(B) = \frac{3}{6}$, joten

$$P(A) + P(B) = \frac{1}{6} + \frac{3}{6} = \frac{4}{6} = \frac{2}{3}.$$

Tässä esimerkissä siis $P(A \cup B) = P(A) + P(B)$.

b) Suotuisten tapahtumien joukko $A = \{1,2,3,4,5,6\}$ eli $A = E$, joten

$$P(A) = P(E) = \frac{6}{6} = 1.$$

Klassisen todennäköisyyskäsitteen käyttöä rajoittaa ennen muuta alkeistapahtumien symmetrisyyden vaatimus, joten sitä voidaan soveltaa lähinnä peliteoriaan. Käytännön ongelmien ratkaisemiseen tarvitaan yleisempää todennäköisyyden käsitettä.

Empiirinen lähestymistapa

Monissa käytännön tilanteissa todennäköisyyttä lähestytään empiirisesti eli kokemusperäisesti. Tällöin tutkitaan aluksi jotakin ilmiötä ja siihen liittyviä tulosmahdollisuuksia. Tulosmahdollisuuksien esiintymisestä muodostetaan frekvenssitaulukkoja.

ESIMERKKI 10

Teollisessa laaduntarkkailussa haluttiin saada selville viallisten tuotteiden suhteellinen osuus tarkkailemalla viiden peräkkäisen päivän tuotantoa. Kunkin päivän tuotannosta otettiin 1 000 kappaleen otos, jolloin viallisten lukumäärät olivat 15, 26, 14, 19 ja 20. Laske kunkin päivän viallisten tuotteiden suhteellinen osuus erikseen sekä yhdistämällä ensin kahden ensimmäisen päivän, sitten kolmen ensimmäisen päivän otokset jne.

Merkitään viallisten tuotteiden määrää eri päivinä m_i:llä ($i = 1,2,3,4,5$) ja otoksen kokoa n:llä. Tehdään taulukko viallisten tuotteiden suhteellisesta osuudesta eri päivinä.

	m_i (kpl)	n (kpl)	Viallisten suhteellinen osuus
1. päivä	15	1 000	$\frac{15}{1\,000} = 0,015$
2. päivä	26	1 000	$\frac{26}{1\,000} = 0,026$
3. päivä	14	1 000	$\frac{14}{1\,000} = 0,014$
4. päivä	19	1 000	$\frac{19}{1\,000} = 0,019$
5. päivä	20	1 000	$\frac{20}{1\,000} = 0,020$

Havaitaan, että viallisten suhteellinen osuus vaihtelee 1,4 %:n ja 2,6 %:n välillä. Tästä voidaan päätellä, että koko tuotannossa on viallisia tuotteita noin 1–3 %. Sen tarkempaa arviota on tällä tavalla vaikea esittää.

Tutkitaan seuraavaksi taulukkoa, jossa otokset on yhdistetty.

	m_i (kpl)	n (kpl)	Viallisten suhteellinen osuus
1. päivä	15	1 000	$\frac{15}{1\,000} = 0,015$
2. päivä	41	2 000	$\frac{41}{2\,000} = 0,021$
3. päivä	55	3 000	$\frac{55}{3\,000} = 0,018$
4. päivä	74	4 000	$\frac{74}{4\,000} = 0,019$
5. päivä	94	5 000	$\frac{94}{5\,000} = 0,019$

Otoksia yhdistettäessä vaihtelu on vähäisempää. Kun otosten yhdistetty koko kasvaa, viallisten tuotteiden suhteellinen osuus vakiintuu vajaaseen kahteen prosenttiin. Voidaan olettaa, että otoksen koon kasvaessa tulosten luotettavuus lisääntyy. Voidaan siis päätellä, että viallisten tuotteiden osuus koko tuotannosta on noin 2 %.

Empiirisen todennäköisyyden käyttö soveltuu hyvin tilanteisiin, joissa seurataan jonkin ilmiön eri tulosmahdollisuuksien suhteellisia osuuksia, kuten viallisten tuotteiden, asiakaskäyntien, tapaturmien ja liikenneonnettomuuksien suhteellisia osuuksia eri aikoina.

Tilastollinen lähestymistapa

Oletetaan, että jokin satunnaisilmiö voi tapahtua mielivaltaisen monta kertaa samanlaisissa olosuhteissa. Edelleen oletetaan, että ilmiöstä voidaan havaita erilaisia tapahtumia. Olkoon A eräs tapahtuma, joka voi tapahtua tai olla tapahtumatta. Jos n kerralla tapahtuma A sattuu $n(\mathrm{A})$ kertaa, tapahtuman A suhteellinen osuus (suhteellinen frekvenssi) on

$$\mathrm{P}_n(\mathrm{A}) = \frac{n(\mathrm{A})}{n}.$$

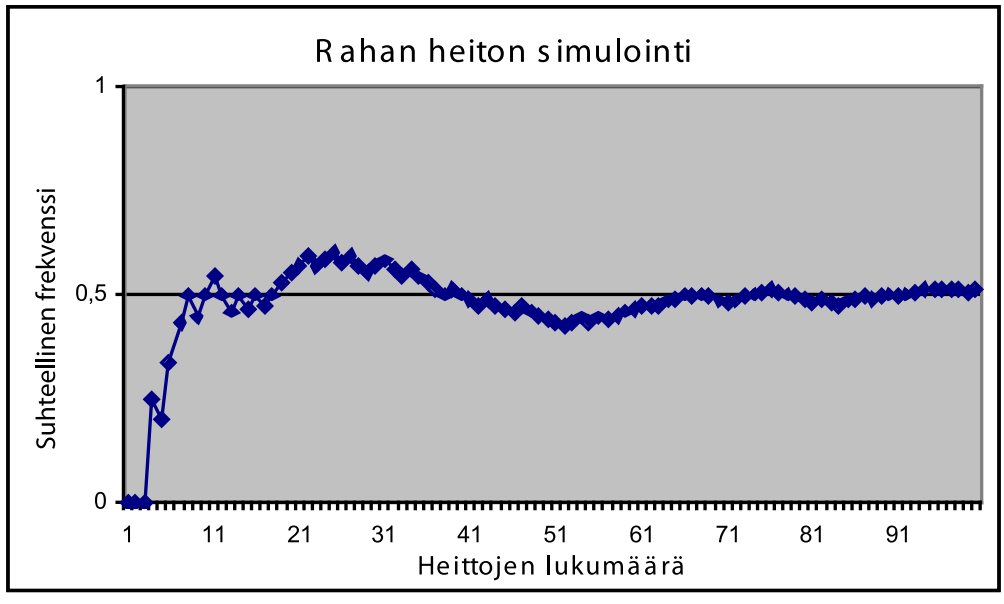
Suhteellisen osuuden raja-arvoa

$$\mathrm{P}(\mathrm{A}) = \lim_{\mathrm{n}\to\infty} \mathrm{P}_n(\mathrm{A}) = \lim_{\mathrm{n}\to\infty} \frac{n(\mathrm{A})}{n}$$

sanotaan tapahtuman A **tilastolliseksi todennäköisyydeksi** (statistical probability).

Tilastollisen todennäköisyyden määritelmä on teoreettinen. Se perustuu **suurten lukujen** lakiin, jonka mukaan toistokokeen tuloksena saatavan tapahtuman suhteellisen osuuden ”tilastollinen” raja-arvo on sama kuin tapahtuman todennäköisyys. Käytännössä suhteellisen frekvenssin tilastollinen stabiliteetti kasvaa, kun koe tai ilmiö toistuu riittävän monta kertaa. Tällä tavoin saadaan riittävän tarkka likiarvo tilastolliselle todennäköisyydelle.

Suhteellisen frekvenssin vakiintumista *n*:n kasvaessa voidaan kuvata simuloimalla rahan heittoa. Kun rahanheittoa simuloitiin Excelin avulla 100 kertaa ja laskettiin kunkin heiton jälkeen kruunujen senhetkinen suhteellinen frekvenssi, saatiin seuraava kuvio:



Aluksi suhteellisen osuuden heilahtelu on satunnaista ja joskus myös voimakasta, mutta se vaimenee vähitellen. Tässäkin tapauksessa kruunujen suhteellinen osuus 100 heiton sarjassa oli lähes täsmälleen 0,5.

Aksiomaattinen lähestymistapa

Aksiomaattinen lähestymistapa perustuu venäläisen Kolmogorovin esittämiin aksioomiin, joilla kuvataan satunnaismallien yleinen rakenne. Keskeiset aksioomat ovat seuraavat:

1. Jokaiselle tapahtumalle A on voimassa
   $0 \leq P(A) \leq 1$.
2. Varman tapahtuman E todennäköisyys on yksi.
   $P(E) = 1$.
3. Jos tapahtumat ovat toisensa poissulkevia, eli ne eivät voi sattua yhtä aikaa,
   $P(A \cup B) = P(A \text{ tai } B) = P(A) + P(B)$.

Kolmogorovin aksioomat ovat sopusoinnussa sekä klassisen että empiirisen lähestymistavan kanssa.

5.3 Todennäköisyyden laskusääntöjä

ESIMERKKI 11

Lastenkutsuille oli hankittu korillinen (24 pulloa) limonadia. Limonadipulloista kolmannes oli punaista limonadia, kolmannes keltaista limonadia, neljäsosa colajuomaa ja loput väritöntä limonadia. Nimipäiväsankari Liisa sai ensimmäisenä valita haluamansa limonadipullon. Koska hän oli epävarma valinnastaan, hän pani silmänsä kiinni ja valitsi satunnaisesti käteensä sattuvan pullon. Mikä on todennäköisyys, että Liisa sai punaista limonadia tai colajuomaa?

Korissa oli 8 punaista, 8 keltaista ja 2 väritöntä limonadipulloa sekä 6 colajuomapulloa.

Merkitään: A = {juoma on punaista}; B = {juoma on colaa}. Tapahtumat ovat selvästi erillisiä. Tällöin todennäköisyyden yhteenlaskuominaisuuden nojalla $P(A \cup B) = P(A) + P(B)$. Koska $n(A) = 8$, $n(B) = 6$ ja $n(E) = 24$, kysytty todennäköisyys on

$$P(A \cup B) = \frac{8}{24} + \frac{6}{24} = \frac{14}{24} = \frac{7}{12} \approx 0{,}583 = 58{,}3\ \%.$$

Edellä oleva voidaan yleistää.

Erillisten tapahtumien yhteenlaskusääntö

Oletetaan, että tapahtumat A_1, A_2,…, A_n ovat erillisiä. Tällöin todennäköisyys, että jokin tapahtumista A_i *(i = 1, 2,..., n)* sattuu, on

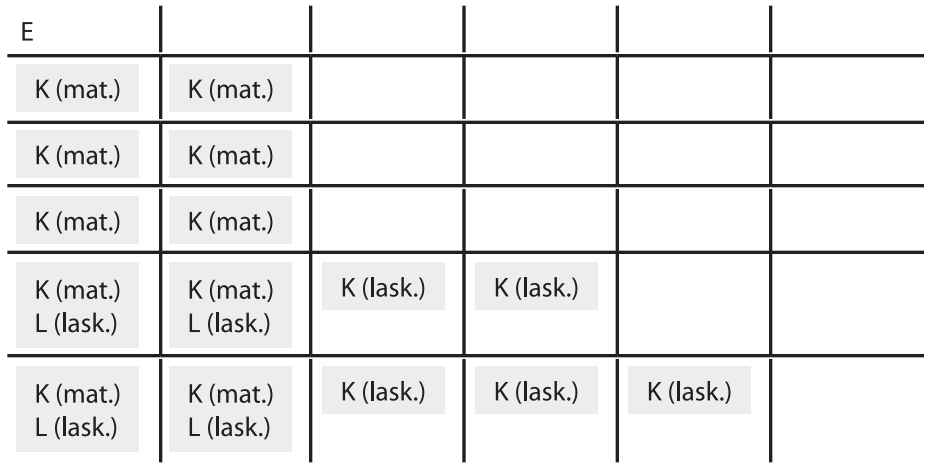
$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n).$$

Yleinen yhteenlaskusääntö on edellistä monimutkaisempi, jos tapahtumat eivät ole erillisiä.

ESIMERKKI 12

Luokalla on 36 opiskelijaa. Heistä kymmenellä on matematiikassa ja yhdeksällä laskentatoimessa kiitettävä arvosana. Neljällä opiskelijalla on kiitettävä arvosana molemmissa oppiaineissa. Työnantaja tarjoaa kesätyöpaikan yhdelle opiskelijalle, jolla on kiitettävä arvosana joko matematiikassa tai laskentatoimessa tai molemmissa oppiaineissa. Mikä on todennäköisyys saada vaatimukset täyttävä opiskelija, jos tämä valitaan arpomalla kaikista luokan opiskelijoista?

Olkoon A = {opiskelijat, joilla on kiitettävä matematiikassa} ja B = {opiskelijat, joilla on kiitettävä laskentatoimessa}. Tällöin $A \cap B$ = {opiskelijat, joilla on kiitettävä molemmissa oppiaineissa} $n(A) = 10$; $n(B) = 9$; $n(A \cap B) = 4$ ja $n(E) = 36$. Esitetään joukot A, B, $A \cap B$ ja E seuraavan kuvion avulla:



Kuviossa yksi ruutu vastaa yhtä opiskelijaa.

Kuvasta käy ilmi, että

$n(A \cup B) = 15$ (varjostetut ruudut yhteensä).

Kysytty todennäköisyys on siten

$$P(A \cup B) = \frac{15}{36} = \frac{5}{12} \approx 0,417.$$

Toisaalta voidaan ajatella, että suotuisten alkeistapahtumien joukko muodostuu opiskelijoista, joilla on kiitettävä arvosana matematiikassa, ja opiskelijoista, joilla on kiitettävä arvosana laskentatoimessa. Näiden yhteenlasketusta määrästä 19 on kuitenkin vähennettävä niiden opiskelijoiden määrä, jotka tulevat lasketuiksi mukaan kahteen kertaan (ne 4 opiskelijaa, joilla on kiitettävä molemmissa oppiaineissa).

$n(A) + n(B) - n(A \cap B) = 10 + 9 - 4 = 15$, joten tässä tapauksessa

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{10}{36} + \frac{9}{36} - \frac{4}{36} = \frac{15}{36} = \frac{5}{12} \approx 0,417.$$

Kahden tapahtuman yleinen yhteenlaskusääntö

Olkoon tapahtuman A todennäköisyys P(A) ja tapahtuman B todennäköisyys P(B). Tällöin todennäköisyys, että joko tapahtuma A tai tapahtuma B sattuu tai molemmat sattuvat, on
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

ESIMERKKI 13

Tilastojen mukaan puolet suomalaisista on ylipainoisia. Valitaan satunnaisesti kaksi suomalaista. Mikä on todennäköisyys, että molemmat valitut ovat ylipainoisia?

Merkitään: 1 = valittu henkilö on ylipainoinen, ja 2 = valittu henkilö ei ole ylipainoinen.

Perusjoukon E muodostavat 4 tulosmahdollisuutta, eli E = {(1, 1), (1, 2), (2, 1), (2, 2)}; $n(E) = 4$. Koska puolet suomalaisista on ja puolet ei ole ylipainoisia, kaikki perusjoukon E tulosmahdollisuudet ovat yhtä todennäköisiä.

Olkoon A = {1. valittu henkilö on ylipainoinen}; B = {2. valittu henkilö on ylipainoinen}. Suotuisten tapahtumien joukko

$A \cap B = \{(1, 1)\}$ eli $n(A \cap B) = 1$.

Kysytty todennäköisyys on $P(A \cap B) = \frac{1}{4} = 0,25$.

Samaan tulokseen päästään kertomalla keskenään tapahtumien A ja B todennäköisyydet:

$$P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0,25.$$

Edellisessä esimerkissä tapahtumat A ja B olivat toisistaan **riippumattomia**. Tapahtuman B sattuminen ei millään tavalla riipu tapahtuman A sattumisesta ja päinvastoin.

Riippumattomien tapahtumien kertolaskusääntö

Olkoot A_1, A_2,... A_n toisistaan riippumattomia tapahtumia, joiden todennäköisyydet ovat $P(A_1)$, $P(A_2)$,..., $P(A_n)$. Tällöin todennäköisyys, että n satunnaisilmiössä tapahtuvat kaikki edellä mainitut tapahtumat, on $P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n)$.

ESIMERKKI 14

Hyvä koripalloilija heittää 80 % vapaaheitoistaan koriin. Mikä on todennäköisyys, että neljä peräkkäistä heittoa menee koriin? Heitot oletetaan toisistaan riippumattomiksi.

Olkoon A_i = {*i*. heitto menee koriin}, $i = 1,2,3,4$.

$P(A_i) = 0{,}8$, joten

$P(A_1 \cap A_2 \cap A_3 \cap A_4) = 0{,}8 \cdot 0{,}8 \cdot 0{,}8 \cdot 0{,}8 = 0{,}8^4 \approx 0{,}41$.

ESIMERKKI 15

Maahantuoja on havainnut, että 15 % sen maahantuomista ilotulitteista on viallisia. Valitaan 20 ilotulitteen erästä satunnaisesti peräkkäin 2 ilotulitetta. Mikä on todennäköisyys, että molemmat valitut ovat viallisia, jos viallisten osuus kyseisessä erässä on maahantuojan havainnon mukainen?

Koska viallisia on 15 %, 20 kappaleen erässä on 3 viallista ($0{,}15 \cdot 20 = 3$).

Merkitään: A = {1. valittu ilotulite on viallinen}; B = {2. valittu ilotulite on viallinen}.

Tapahtuman A todennäköisyys $P(A) = \frac{3}{20}$.

Tapahtuman B todennäköisyys riippuu siitä, oliko ensimmäinen valittu viallinen vai virheetön. Jos ensimmäinen oli viallinen, tapahtuman B todennäköisyys on $\frac{2}{19}$.

Tällaista todennäköisyyttä sanotaan **ehdolliseksi** todennäköisyydeksi ja sitä merkitään $P(B|A)$:

$$P(A \cap B) = P(A) \cdot P(B|A) = \frac{3}{20} \cdot \frac{2}{19} = \frac{3}{190} \approx 0{,}0158.$$

Yleinen kertolaskusääntö

Olkoot A ja B tapahtumia, joiden todennäköisyydet ovat P(A) ja P(B). Tällöin todennäköisyys, että tapahtumat A ja B sattuvat peräkkäin, on $P(A \cap B) = P(A) \cdot P(B|A)$.

Yleisesti on voimassa, että

$P(A_1 \cap A_2 \cap \ldots \cap A_n) =$

$P(A_1) \cdot P(A_2|A_1) \cdot \ldots \cdot P(A_n \mid A_1, A_2, \ldots, A_{n-1})$.

Huomautus

Merkintä P(B|A) luetaan: ”Todennäköisyys, että sattuu tapahtuma B ehdolla, että tapahtuma A on sattunut.”

ESIMERKKI 16

Laatikossa on 500 arpaa, joista voittoarpoja on 40. Nostetaan peräkkäin kolme arpaa. Mikä on todennäköisyys, ettei mikään niistä ole voittoarpa?

A_i = {*i.* arpa ei ole voittoarpa}, $i = 1,2,3$.

$P(A_1) = \frac{460}{500}$, $P(A_2|A_1) = \frac{459}{499}$ ja $P(A_3|A_1,A_2) = \frac{458}{498}$

Todennäköisyys, että mikään kolmesta arvasta ei ole voittoarpa, on

$$P(A_1 \cap A_2 \cap A_3) = \frac{460}{500} \cdot \frac{459}{499} \cdot \frac{458}{498} \approx 0{,}778 .$$

Komplementti

Kun rahaa heitetään kerran, saadaan aina joko kruunu tai klaava. Toisaalta molempia ei koskaan saada samalla kertaa. Tällaisia tapahtumia, jotka eivät voi tapahtua yhtä aikaa mutta joista jompikumpi tapahtuu aina, sanotaan toistensa **komplementtitapahtumiksi**.

Olkoot A ja B perusjoukon E tapahtumia. Jos

1. A ja B ovat erillisiä eli $A \cap B = \Phi$
2. A ja B yhdessä käsittävät kaikki E:n tapahtumat eli $A \cup B = E$,

sanotaan, että A ja B ovat toistensa komplementtitapahtumia.

Tapahtuman A komplementtitapahtumaa merkitään $\overline{A}$:lla.

Koska $A \cup \overline{A} = E$ ja tapahtumat A ja $\overline{A}$ ovat erillisiä, on $P(A \cup \overline{A}) = P(A) + P(\overline{A}) = P(E) = 1$. Tästä seuraa komplementtitapahtuman todennäköisyyttä koskeva sääntö.

Komplementtitapahtuman todennäköisyys

Olkoot A ja $\overline{A}$ toistensa komplementtitapahtumia ja olkoon tapahtuman A todennäköisyys P(A). Tällöin komplementtitapahtuman todennäköisyys

$P(\overline{A}) = 1 - P(A)$.

ESIMERKKI 17

Mikä on esimerkin 16 tapauksessa todennäköisyys, että saadaan vähintään yksi voittoarpa?

Olkoon tapahtuma A = {mikään arpa ei voita}. Tapahtuman A todennäköisyys on esimerkin 16 mukaan 0,778. Tämän komplementtitapahtuma on, että saadaan vähintään yksi voittoarpa.

$P(\overline{A}) = 1 - P(A) = 1 - 0,778 = 0,222$

TEHTÄVIÄ

1. Yrityksen johtajan on asetettava kuusi työpaikan hakijaa paremmuusjärjestykseen. Kuinka monta erilaista mahdollisuutta on olemassa?

2. Opiskelija tekee opinnäytetyön, johon liittyy otokseen valittujen opiskelijoiden haastattelu. Kuinka monessa eri järjestyksessä hän voi haastatella otokseen valittuja, jos otoksen koko on
a) 8 opiskelijaa,
b) 12 opiskelijaa,
c) 15 opiskelijaa?

3. Ravintolan ruokalistalla on 3 erilaista alkuruokaa, 6 liharuokaa, 4 kalaruokaa ja 5 jälkiruokaa. Asiakas syö aterian, johon kuuluu alkuruoka, liha- tai kalaruoka ja jälkiruoka. Kuinka monta erilaista ateriavaihtoehtoa hänellä on valittavanaan?

4. Kuinka monta sellaista nelinumeroista lukua on olemassa, joissa sama numero esiintyy vain kerran?

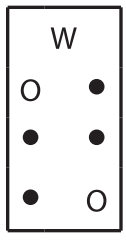
5. Kutsuilla on 10 avioparia. Jokainen läsnäolija kättelee jokaista muuta paitsi aviopuolisoaan. Kuinka monta kättelyä suoritetaan? (Yo 1998 K, lyhyt)

6. Tentissä on 10 kysymystä, joista enintään kuuteen saa vastata. Opiskelija on valmistautunut hyvin ja osaisi vastata jokaiseen kysymykseen. Hän harkitsee, mitkä 6 kysymystä hän valitsee. Kuinka monta vaihtoehtoa hänellä on?

7. Kokelaan ylioppilastutkinnon matematiikan koe muodostuu 10 tehtävästä. Vanhanmuotoisessa matematiikan kokeessa oli 10 tehtävää ja osassa tehtävistä oli kaksi vaihtoehtoa, joista sai laskea vain toisen; nykyisessä kokeessa saa 15 tehtävästä vapaasti valita suoritettavakseen 10 tehtävää. Kuinka monta erilaista 10 tehtävän kokonaisuutta on
a) vanhanmuotoisessa kokeessa, jossa viidessä tehtävässä on kaksi vaihtoehtoa, ja
b) nykyisessä uudenmuotoisessa kokeessa?
(Yo 2000 S, lyhyt)

8. Veikkaaja aavistaa, että seuraavan viikon oikeassa veikkausrivissä on 6 ykköstä, 3 ristiä ja 4 kakkosta. Kuinka monta riviä hänen on veikattava saadakseen varmasti täysosuman, jos aavistus pitää paikkaansa?

9. Ranskalaisen Louis Braillen vuonna 1825 kehittämä pistekirjoitus on kohokirjoitusta, jota luetaan sormin. Pistekirjoitusjärjestelmässä kutakin merkkiä kohti on käytettävissä kuusi kiinteää paikkaa, joihin voidaan asettaa yhdestä kuuteen pistettä. (Esimerkkinä kuviossa on kirjain W.) Kuinka monta erilaista merkkiä järjestelmässä voidaan esittää?
(Yo 2007 K, lyhyt)



10. Suomalaisessa lotossa arvotaan 7 numeroa 39 numerosta. Yhteispohjoismaisessa Viking-lotossa puolestaan arvotaan 6 numeroa 48 numerosta. Kummassa lotossa on suurempi mahdollisuus saada täysosuma, kun täytetään yksi ruudukko?

11. Henkilö lottoaa suomalaista lottoa 10 rastin järjestelmällä. Kuinka paljon järjestelmän käyttö maksaa, kun yhden ruudukon hinta on 0,80 euroa?

12. 52 kortin korttipakasta otetaan 5 korttia. Kuinka monta erilaista 5 kortin ”kättä” voidaan saada?

13. Laaduntarkkailussa otetaan 50 kappaleen valmistuserästä 3 kappaleen otos. Kuinka monta erilaista otosta on olemassa?

14. Korttipakasta vedetään kortti. Mikä on todennäköisyys, että saadaan
a) hertta,
b) ruutuässä tai patakuningas?

15. Luokalla on 12 poikaa ja 20 tyttöä. Heistä yksi arvotaan edustamaan luokkaa eräässä tilaisuudessa. Millä todennäköisyydellä joku pojista tulee valituksi?

16. Laatikossa on 10 punaista, 7 keltaista ja 6 valkoista palloa. Nostetaan satunnaisesti yksi pallo laatikosta. Mikä on todennäköisyys, että se on
a) punainen tai valkoinen,
b) keltainen tai valkoinen?

17. Heitetään yhtä aikaa kahta noppaa. Mikä on todennäköisyys, että silmälukujen summa on 9 tai 10?

18. Noppaa heitetään kaksi kertaa. Mikä on todennäköisyys, että jälkimmäisellä heitolla saadaan suurempi silmäluku kuin edellisellä? Jos noppaa heitetään kolme kertaa ja ensimmäisellä heitolla on saatu kolmonen, mikä on todennäköisyys, että toisella heitolla saadaan suurempi silmäluku kuin ensimmäisellä ja kolmannella heitolla suurempi kuin toisella? (Yo 2002 K, lyhyt)

19. Heitetään yhtä aikaa kahta rahaa. Mikä on todennäköisyys, että saadaan
a) kruunu ja klaava,
b) kaksi klaavaa?

20. Erään kunnan kunnallisvaaleissa viisi puoluetta oli asettanut ehdokkaita. Äänimäärät jakautuivat seuraavasti:

Puolue	A	B	C	D	E	Ei äänestänyt
Äänimäärä	1 810	790	3 061	1 514	2 357	1 858

Mikä on todennäköisyys, että mielivaltaisesti valittu näissä vaaleissa äänioikeutettu kuntalainen
a) äänesti puoluetta C,
b) jätti äänestämättä,
c) äänesti puoluetta A tai jätti äänestämättä?

21. Teollisuuslaitoksen tuotantolinjalla suoritettiin säännöllisin välein laaduntarkkailua. Aiemmin oli havaittu, että virheettömien tuotteiden suhteellinen osuus oli 0,995. Erään päivän tuotanto oli 7 516 kappaletta, joista viallisia oli 46 kappaletta. Oliko kyseisen päivän tuotanto laadultaan keskimääräistä parempaa vai huonompaa?

22. Laaduntarkkailussa pyrittiin selvittämään viallisten tuotteiden suhteellinen osuus tarkkailemalla kuuden peräkkäisen päivän tuotantoa. Kunkin päivän tuotannosta poimitut otokset ja viallisten tuotteiden lukumäärät esitetään seuraavassa taulukossa.

Otoskoko	163	269	244	190	413	222
Viallisten määrä	4	18	9	5	19	10

Päättele, kuinka monta prosenttia koko tuotannosta on viallisia.

23. Luokalla on 32 opiskelijaa. Heistä 14 on englanninkielentaitoisia ja 8 saksankielentaitoisia. Molempia kieliä osaa 3 opiskelijaa. Mikä on todennäköisyys, että arpomalla valittu opiskelija osaa ainakin toista kyseisistä kielistä?

24. Mikä on todennäköisyys, että lottoarvonnassa (39 palloa mukana arvonnassa) ensimmäisen arvotun pallon numero on
a) 5 tai parillinen,
b) kolmella tai neljällä jaollinen?

25. Kunnan henkikirjoitetun väestön ikäjakauma oli seuraava:

Ikä (vuosia)	Naisia	Miehiä
0–17	312	298
18–39	267	283
40–55	241	193
56–	183	116

Valitaan satunnaisesti yksi kunnassa henkikirjoitettu henkilö. Mikä on todennäköisyys, että hän on

a) alle 18-vuotias tai vähintään 56-vuotias,

b) nainen tai 40–55-vuotias?

26. Tuotteista 5 % on viallisia. Suuresta tuote-erästä otetaan

a) 2 kappaleen,

b) 10 kappaleen otos.

Mikä on todennäköisyys, että koko otos on virheetön?

27. Suomessa syntyvistä lapsista 51,3 % on poikia. Kuinka suurella todennäköisyydellä seitsenlapsisen Jukolan perheen kaikki lapset ovat poikia? (Yo 1993 K, lyhyt)

28. Tilastojen mukaan 71 % 25-vuotiaista miehistä ja 82 % samanikäisistä naisista elää vielä 60-vuotiaina. Satunnaisesti toisensa valinneet 25-vuotiaat mies ja nainen menevät naimisiin. Mikä on todennäköisyys, että

a) molemmat elävät 60-vuotiaina

b) vain nainen elää 60-vuotiaana

c) vain mies elää 60-vuotiaana

d) molemmat ovat kuolleet ennen 60-vuotispäiväänsä?

29. Jokeripelissä arvotaan peräkkäin seitsemän numeroa numeroista 0–9 toisistaan riippumatta. Mikä on todennäköisyys, että

a) numeroiden muodostama luku on 1122333

b) kaikki arvotut numerot ovat samoja?

30. Vakioveikkauksessa on 13 kohdetta. Laske todennäköisyys, että satunnaisesti valitussa rivissä on

a) täysosuma,

b) 12 oikein?

31. 50 kappaleen tuote-erässä on 8 virheellistä. Tuote-erästä otetaan peräkkäin kahden kappaleen satunnaisotos. Mikä on todennäköisyys, että

a) molemmat ovat virheettömiä,

b) jompikumpi on virheetön?

32. Yrityksen johtoryhmä muodostuu 8 henkilöstä, joista 3 on naisia. Johtoryhmästä valitaan peräkkäin arpomalla kolmen henkilön toimikunta valmistelemaan erästä asiaa. Mikä on todennäköisyys, että
a) kaikki valitut ovat miehiä
b) kaikki valitut ovat naisia
c) kaksi ensin valittua ovat miehiä ja kolmas valittu on nainen?

33. Olkoon tarkasteltava satunnaisilmiö nopan heitto. Mikä on tapahtuman A = {silmäluku on suurempi kuin 2} komplementtitapahtuma?

34. Arpajaisissa on 1 200 arpaa, joista 5 % on voittoarpoja. Henkilö ostaa kolme arpaa. Mikä on todennäköisyys, että
a) hän ei saa yhtään voittoarpaa
b) ensimmäinen arpa on voittoarpa mutta muut eivät ole?

35. Millä todennäköisyydellä edellisen tehtävän henkilö saa vähintään yhden voittoarvan?

36. Potilaista 15 % on allergisia sulfalle. Valitaan satunnaisesti 5 potilasta. Mikä on todennäköisyys, että
a) kukaan heistä ei ole allerginen sulfalle,
b) vähintään yksi heistä on allerginen sulfalle?

37. Heitetään noppaa kuusi kertaa peräkkäin. Millä todennäköisyydellä saadaan vähintään kerran silmäluku 6?

38. Laaduntarkkailussa on todettu, että lopputarkastukseen tulevista tuotteista 3,5 % on viallisia. Otetaan tarkastukseen tulevasta suuresta tuote-erästä 5 kappaleen satunnaisotos. Mikä on todennäköisyys, että
a) vähintään yksi on viallinen,
b) enintään neljä on viallisia?

39. Todennäköisyys, ettei satunnaisesti valitussa ydinvoimalassa tapahdu sen käyttöiän aikana vakavaa onnettomuutta, on 99,99 %. Mikä on todennäköisyys, että vähintään yhdessä ydinvoimalassa sattuu vakava onnettomuus, jos maapallolla on
a) 260,
b) 500 ydinvoimalaa?

40. Makeisia valmistava yritys on todennut, että lopputarkastukseen tulevista tuotteista kahdessa prosentissa on raaka-ainevirhe. Lisäksi kolmessa prosentissa tuotteista on valmistusvirhe ja neljässä prosentissa pakkausvirhe. Millä todennäköisyydellä lopputarkastukseen tulevassa tuotteessa on
a) vähintään yksi virhe,
b) enintään kaksi virhettä?

6
Todennäköisyysjakaumat

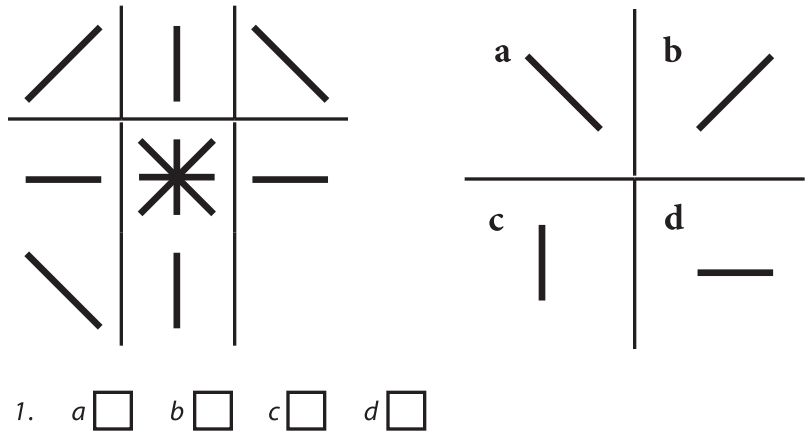
Tässä luvussa käsiteltävät todennäköisyysjakaumat ovat tilastotieteen keskeisen osa-alueen, tilastollisen päättelyn, perusta. Edellisessä luvussa käsiteltyjen todennäköisyyksien laskusääntöjen avulla voidaan laskea yksittäisten tapahtumien todennäköisyyksiä. Todennäköisyysjakaumia kuvaavien mallien avulla hallitaan kaikki johonkin ilmiöön liittyvät todennäköisyydet.

Suomessakin toimii kansainvälinen huippuälykkäiden yhdistys Mensa ry. Sen jäseneksi pääsee vähintään 16-vuotias henkilö, joka saavuttaa Mensan käyttämässä Cattell & Cattellin kuviopäättelytestissä vähintään pistemäärän 148. Testi ei ole kulttuurisidonnainen, eikä siinä ole lainkaan sanallisia tai matemaattisia tehtäviä.

Jos kaikki suomalaiset 16 vuotta täyttäneet testattaisiin kyseisellä testillä, testipistemäärien keskiarvo olisi 100 ja keskihajonta 24. Mensan pääsyvaatimus on siten keskimääräinen älykkyys + 2 kertaa keskihajonta.

Kuinka moni neljästä miljoonasta 16 vuotta täyttäneestä suomalaisesta pääsisi älykkyytensä puolesta Mensa ry:n jäseneksi? Kyseessä on ongelma, joka voidaan ratkaista **normaalijakauman** avulla.

Täytä vasemmanpuoleisessa kuviossa oleva tyhjä kenttä siihen loogisesti kuuluvalla kuviolla a, b, c tai d.



1. a ☐ b ☐ c ☐ d ☐

Tyypillinen älykkyystestin tehtävä.

6.1 Diskreetit todennäköisyysjakaumat

Klassisella todennäköisyyslaskennalla on merkitystä lähinnä arpajais- ja peliteoriassa. Empiiristä ja tilastollista todennäköisyyttä voidaan käyttää esimerkiksi satunnaisilmiön yksittäiseen tapahtumaan liittyvän riskin arvioinnissa, mutta kaikkien riskien hallinta edellyttää todennäköisyysmallien tuntemista.

Ryhdytään tarkastelemaan, millä tavalla reaalimaailmassa havaittavia satunnaisilmiöitä voidaan kuvata **todennäköisyysmallien** avulla.

Satunnaismuuttuja

Muuttujaa X sanotaan **satunnaismuuttujaksi** (random variable), jos se liittyy johonkin satunnaisilmiöön ja on täysin määrätty heti, kun koe on tehty tai ilmiö on tapahtunut. Satunnaismuuttujia merkitään isoilla kirjaimilla X, Y, Z,... ja niiden yksittäisiä arvoja pienillä kirjaimilla x, y, z,...

Luvussa 1 empiiriset muuttujat jaettiin diskreetteihin ja jatkuviin muuttujiin. Aivan samalla tavalla satunnaismuuttujat voivat olla diskreettejä tai jatkuvia. Satunnaismuuttuja on **epäjatkuva** eli **diskreetti,** jos se voi saada vain tiettyjä arvoja. Vastaavasti **jatkuva** satunnaismuuttuja voi tietyllä välillä saada minkä arvon tahansa.

ESIMERKKI 1

Simuloidaan kahden rahan yhtä aikaa tapahtuvaa heittoa ”heittämällä” niitä tietokoneohjelman avulla yhteensä 100 kertaa. Kruunujen (kr) määrä simulaatiossa jakaantui seuraavasti:

Kruunuja	Lukumäärä
0	23
1	48
2	29

Tässä on kyseessä empiirinen jakauma.

Kahta rahaa heitettäessä tulosmahdollisuudet ovat {(kr, kr); (kr, kl); (kl, kr); (kl, kl)}. Ilmiselvästi jokainen tulosmahdollisuus on yhtä todennäköinen. Voidaan siis päätellä, että kruunujen määrän teoreettinen jakauma on seuraava:

Kruunuja	Todennäköisyys
0	25 %
1	50 %
2	25 %

Tällaista jakaumaa sanotaan **todennäköisyysjakaumaksi**. Sen avulla voidaan ennustaa kruunujen lukumäärää, kun kahta rahaa heitetään yhtä aikaa n kertaa.

Diskreetti todennäköisyysjakauma

Satunnaismuuttujan X arvot x_i ja niihin liittyvät todennäköisyydet p_i muodostavat diskreetin todennäköisyysjakauman, joka voidaan esittää seuraavasti:

X	p
x_1	p_1
x_2	p_2
…	…
x_n	p_n

ESIMERKKI 2

Heitetään yhtä aikaa kahta noppaa. Määritä silmälukujen summan todennäköisyysjakauma.

Tulosmahdollisuuksia on yhteensä 36 kappaletta:

2. noppa / 1. noppa	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Silmäluku 2 saadaan vain, jos molempien noppien silmäluku on 1. Satunnaismuuttujan arvoa 2 vastaa siten todennäköisyys 1/36. Suurin todennäköisyys on saada summaksi 7: {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}. Sen todennäköisyys on 6/36.

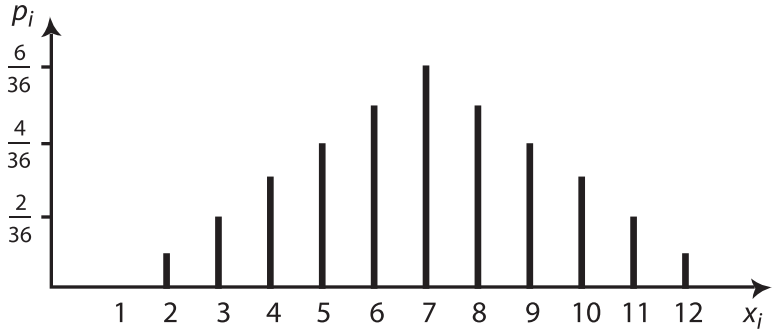
Todennäköisyysjakaumaksi saadaan seuraava taulukko:

Silmälukujen summa	Todennäköisyys	Kertymätodennäköisyys
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	36/36

Kertymätodennäköisyys (cumulative probability) määritetään samalla tavalla kuin empiirisissä jakaumissa kertymäfrekvenssi.

Edellä saadut tulokset merkitsevät, että kun heitetään kahta noppaa, kannattaa ennen heittoa arvata silmälukujen summaksi 7.

Todennäköisyysjakauma voidaan esittää graafisesti seuraavan kuvion avulla:



Diskreetin jakauman kertymäfunktio

Oletetaan, että satunnaismuuttuja X saa arvot x_1, x_2, x_3,… todennäköisyyksin p_1, p_2, p_3,… Tällöin satunnaismuuttujan jakaumaan liittyy kertymäfunktio

$$\mathrm{F}(x) = \mathrm{P}(X \le x) = \sum_i p_i \,.$$

Indeksi i saa kaikki arvot, joilla $x_i \le x$. Kertymäfunktio $\mathrm{F}(x)$ osoittaa siis todennäköisyyden, jolla satunnaismuuttuja X saa enintään arvon x.

ESIMERKKI 3

Mikä on todennäköisyys, että kahta noppaa heitettäessä silmälukujen summa on a) 6, 7 tai 8, b) enintään 6, c) vähintään 10?

a) Lasketaan yhteen X:n arvoja 6, 7 ja 8 vastaavat todennäköisyydet.

$$\mathrm{P}(6 \le X \le 8) = \frac{5}{36} + \frac{6}{36} + \frac{5}{36} = \frac{16}{36} = \frac{4}{9} \approx 0,444$$

b) $$\mathrm{P}(X \le 6) = \mathrm{P}(X = 2) + \mathrm{P}(X = 3) + \mathrm{P}(X = 4) + \mathrm{P}(X = 5) + \mathrm{P}(X = 6)$$
$$= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} = \frac{15}{36} = \frac{5}{12} \approx 0,417$$

c) $$\mathrm{P}(X \ge 10) = \mathrm{P}(X = 10) + \mathrm{P}(X = 11) + \mathrm{P}(X = 12)$$
$$= \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{6}{36} = \frac{1}{6} \approx 0,167$$

Edellisessä esimerkissä kysytyt todennäköisyydet saatiin laskemalla yhteen yksittäisten tapahtumien todennäköisyydet eli **pistetodennäköisyydet**. Kertymäfunktion avulla kysytyt todennäköisyydet voidaan määrittää edellistä vaivattomammin.

ESIMERKKI 4

Määritä kahden nopan heittoon liittyvän todennäköisyysjakauman kertymäfunktio.

$F(X) = 0$, kun $x < 2$
(Silmälukujen summa ei voi olla pienempi kuin 2.)

$F(x) = \frac{1}{36}$, kun $2 \le x < 3$

$F(x) = \frac{3}{36}$, kun $3 \le x < 4$

$F(x) = \frac{6}{36}$, kun $4 \le x < 5$

$F(x) = \frac{10}{36}$, kun $5 \le x < 6$

$F(x) = \frac{15}{36}$, kun $6 \le x < 7$

$F(x) = \frac{21}{36}$, kun $7 \le x < 8$

$F(x) = \frac{26}{36}$, kun $8 \le x < 9$

$F(x) = \frac{30}{36}$, kun $9 \le x < 10$

$F(x) = \frac{33}{36}$, kun $10 \le x < 11$

$F(x) = \frac{35}{36}$, kun $11 \le x < 12$

$F(x) = 1$, kun $x \ge 12$

Diskreetin jakauman kertymäfunktion ominaisuuksia

1. $P(X = x_i) = F(x_i) - F(x_{i-1})$
2. $P(x_i < X \leq x_j) = F(x_j) - F(x_i)$

ESIMERKKI 5

Laske esimerkki 3 käyttämällä hyväksi kertymäfunktiota.

a) $P(6 \leq X \leq 8) = P(5 < X \leq 8) = F(8) - F(5) =$

$$\frac{13}{18} - \frac{5}{18} = \frac{8}{18} = \frac{4}{9} \approx 0,444$$

b) $P(X \leq 6) = F(6) = \frac{5}{12} \approx 0,417$

c) Tapahtuman $P(X \geq 10)$ komplementtitapahtuma on $P(X < 10) = P(X \leq 9)$.

Siten $P(X \geq 10) = 1 - P(X \leq 9) = 1 - F(9) = 1 - \frac{30}{36} = \frac{6}{36} = \frac{1}{6} \approx 0,167.$

Todennäköisyysjakauman kuvaamiseen käytetään erilaisia tunnuslukuja, joista tärkeimmät ovat **odotusarvo** ja **keskihajonta**. Näistä edellinen vastaa empiirisen jakauman keskiarvoa ja jälkimmäinen keskihajontaa. Odotusarvo tarkoittaa tapahtumaa, joka sattuu useimmin, jos satunnaisilmiö toistuu riittävän monta kertaa eli se on satunnaisilmiön todennäköisin tulos. Keskihajonta puolestaan osoittaa, kuinka laajalle tapahtuman todennäköisyydet ovat levinneet odotusarvon ympärille.

Diskreetin jakauman odotusarvo

Oletetaan, että satunnaismuuttuja X saa arvot $x_1, x_2, x_3\ldots$ todennäköisyyksin $p_1, p_2, p_3\ldots$ Satunnaismuuttujan X odotusarvo eli keskiarvo (mean)

$$EX = \mu = \sum_i p_i x_i.$$

ESIMERKKI 6

Kahden nopan heitossa saatavan silmälukujen summan odotusarvo on

$$EX = \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \frac{3}{36} \cdot 4 + \ldots + \frac{2}{36} \cdot 11 + \frac{1}{36} \cdot 12 = 7.$$

Tulos voidaan tulkita niin, että kun heitetään yhtä aikaa kahta noppaa esimerkiksi 1 000 kertaa, silmälukujen summan keskiarvo on lähellä lukua 7. Toisaalta, jos keskiarvo poikkeaa huomattavasti luvusta 7, voidaan arvioida, että nopat ovat ”virheellisiä”.

ESIMERKKI 7

Metallialan yritys aikoo leasingvuokrata automaattisorvin. Vuokra-aika on 10 vuotta ja vuosivuokra 50 000 euroa. Sorvin vuokrauksesta saatavat nettotuotot riippuvat suhdanteista siten, että korkeasuhdanteen aikana ne ovat arvioiden mukaan 90 000 euroa vuodessa ja matalasuhdanteen aikana 60 000 euroa vuodessa. Tilastojen perusteella voidaan arvioida, että 40 % vuosista on sellaisia, jolloin vallitsee korkeasuhdanne, ja 60 % vuosista sellaisia, jolloin on matalasuhdanne. Kannattaako leasingvuokraus, kun vuotuisten tuottojen pitää olla keskimäärin 30 % suurempia kuin vuotuisten kustannusten?

Olkoon X vuosituotto. Lasketaan sen odotusarvo. Jos vuosituoton odotusarvo on vähintään 30 % suurempi kuin vuotuiset kustannukset 50 000 € (1,3 · 50 000 € = 65 000 €), vuokraus on kannattavaa.

Satunnaismuuttujan X jakauma on

x_i	90 000 €	60 000 €
p_i	0,4	0,6

$EX = 0{,}4 \cdot 90\,000\,€ + 0{,}6 \cdot 60\,000\,€ = 72\,000\,€ > 65\,000\,€.$

Vuokraus on kannattavaa.

Keskihajonta

Oletetaan, että satunnaismuuttuja X saa arvot x_1, x_2, x_3,… todennäköisyyksin p_1, p_2, p_3,... Satunnaismuuttujan X keskihajonta (standard deviation) on

$$\mathrm{D}X = \sigma = \sqrt{\sum_i p_i \cdot (x_i - \mu)^2}.$$

Keskihajonnan avulla kuvataan, miten muuttujan arvot ovat jakautuneet keskiarvon ympärille.

Keskihajonnan toinen potenssi $\mathrm{D}^2X = \sigma^2$ on **varianssi**.

ESIMERKKI 8

Kahden nopan heitossa saatavan silmälukujen summan varianssi on

$$\mathrm{D}^2X = \frac{1}{36}\cdot(2-7)^2 + \frac{2}{36}\cdot(3-7)^2 + \frac{3}{36}\cdot(4-7)^2 + \ldots$$

$$+ \frac{2}{36}\cdot(11-7)^2 + \frac{1}{36}\cdot(12-7)^2 \approx 5,833,$$

jolloin keskihajonta $\mathrm{D}X = \sqrt{5,833} \approx 2,42$.

Binomijakauma

Diskreeteistä todennäköisyysjakaumista yleisin on **binomijakauma** (binomial distribution). Se liittyy satunnaisilmiöön, jolla on vain kaksi tulosmahdollisuutta. Tulosmahdollisuudet voivat olla esimerkiksi kyllä–ei, oikein–väärin ja tapahtuu–ei tapahdu.

ESIMERKKI 9

Maanlaajuisissa arpajaisissa 15 % arvoista on voittoarpoja. Ostetaan kolme arpaa. Laske todennäköisyydet, että voittoarpoja on 0, 1, 2 ja 3.

Voittoarpojen lukumäärä X on satunnaismuuttuja, joka voi saada arvon 0, 1, 2 tai 3. Jos voittoarvan saamistodennäköisyyttä merkitään symbolilla p ja voitottoman arvan saamistodennäköisyyttä symbolilla q, on $p = 0,15$ ja $q = 1 - p = 0,85$. Todennäköisyys, että $X = 0$, on

$$\mathrm{P}(X = 0) = qqq = q^3 = 0,85^3 = 0,6141.$$

Todennäköisyys, että saadaan yksi voittoarpa, muodostuu kolmesta eri tulosmahdollisuudesta, koska voittoarpa voi olla ensimmäisenä, toisena tai kolmantena.

$P(X = 1) = pqq + qpq + qqp = C_3^1 \cdot pq^2 = 3 \cdot 0{,}15 \cdot 0{,}85^2 = 0{,}3251$

Vastaavasti kahden voittoarvan saamistodennäköisyys on

$P(X = 2) = ppq + pqp + qpp = C_3^2 \cdot p^2q = 3 \cdot 0{,}15^2 \cdot 0{,}85 = 0{,}0574.$

Todennäköisyys, että kaikki kolme ovat voittoarpoja, on

$P(X = 3) = ppp + p^3 = 0{,}15^3 = 0{,}0034.$

Satunnaismuuttujan X todennäköisyysjakauma on seuraava:

x_i	0	1	2	3
p_i	0,6141	0,3251	0,0574	0,0034

Todennäköisyys, että saadaan kaksi tai kolme voittoarpaa, on pieni.

Käytännössä binomijakaumaan liittyvät todennäköisyydet määritetään **todennäköisyysfunktion** (probability function) avulla.

Binomijakauman todennäköisyysfunktio

Tarkastellaan satunnaisilmiötä, jolla on tulosmahdollisuudet A ja $\overline{A}$. Olkoon A:n todennäköisyys P(A) = p ja $\overline{A}$:n todennäköisyys P($\overline{A}$) = $q = 1 - p$. Tällöin todennäköisyys, että ilmiön tapahtuessa n kertaa tapahtuma A sattuu r kertaa, on

$P(X = r) = p(r) = C_n^r \cdot p^r \cdot q^{n-r}$, jossa $r = 0, 1, 2, \ldots, n$.

Binomijakauman todennäköisyysfunktion $p(r)$ arvot (binomitodennäköisyydet) ovat täysin määrätyt heti, kun toistojen lukumäärä n ja todennäköisyys p tunnetaan. Lukuja n ja p sanotaan binomijakauman **parametreiksi**. Jos satunnaismuuttuja X noudattaa binomijakaumaa parametrein n ja p, sitä merkitään: $X \sim \text{Bin}(n, p)$.

ESIMERKKI 10

Lasitehtaan tuotteista 10 % on viallisia. Otetaan neljän kappaleen satunnaisotos. Mikä on todennäköisyys, että saadaan yksi viallinen?

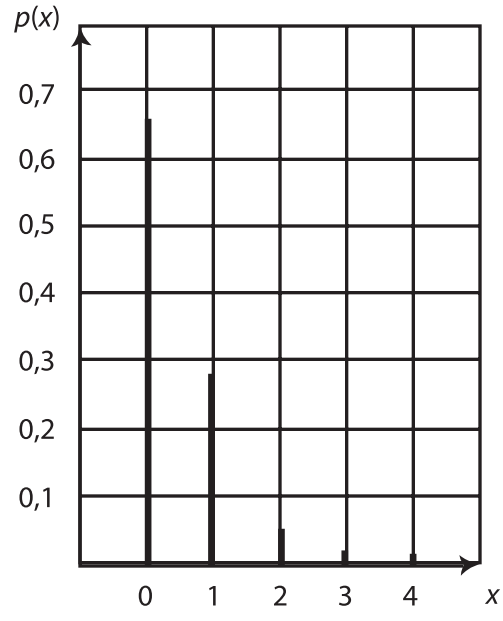
Olkoon X viallisten tuotteiden lukumäärä. Koska $n = 4$ ja $p = 0{,}1$, on $X \sim \text{Bin}(4; 0{,}1)$. Kysytty todennäköisyys on

$p(1) = C_4^1 \cdot 0{,}1^1 \cdot 0{,}9^3 \approx 0{,}292.$

ESIMERKKI 11

Muodosta edellisen esimerkin satunnaismuuttujan todennäköisyysjakauma ja esitä se graafisesti.

r	$p(r) = C_n^r \cdot p^r \cdot q^{n-r}$
0	$p(0) = C_4^0 \cdot 0{,}1^0 \cdot 0{,}9^4 = 0{,}6561$
1	$p(1) = C_4^1 \cdot 0{,}1^1 \cdot 0{,}9^3 = 0{,}2916$
2	$p(2) = C_4^2 \cdot 0{,}1^2 \cdot 0{,}9^2 = 0{,}0486$
3	$p(3) = C_4^3 \cdot 0{,}1^3 \cdot 0{,}9^1 = 0{,}0036$
4	$p(4) = C_4^4 \cdot 0{,}1^4 \cdot 0{,}9^0 = 0{,}0001$



Todennäköisyydet, että otoksessa on enemmän kuin yksi viallinen tuote, ovat hyvin pieniä. Todennäköisyysjakauma on vino.

Binomijakauman vinous vähenee ja symmetrisyys lisääntyy sitä enemmän, mitä lähempänä p on lukua 0,5. Myös parametrin n arvo vaikuttaa binomijakauman symmetrisyyteen siten, että n:n kasvaessa symmetrisyys lisääntyy.

ESIMERKKI 12

Mikä on todennäköisyys, että esimerkin 10 otoksessa on a) enintään yksi viallinen, b) vähintään kaksi viallista?

a) $P(X \leq 1) = P(X = 0) + P(X = 1) = 0{,}6561 + 0{,}2916 = 0{,}9477$

b) $P(X \geq 2) = 1 - P(X < 2) = 1 - P(X \leq 1) = 1 - 0{,}9477 = 0{,}0523$

Binomijakauman kertymäfunktio

Oletetaan, että satunnaismuuttuja $X \sim \text{Bin}(n, p)$. Tällöin satunnaismuuttujan X kertymäfunktio (cumulative distribution function) on

$$\text{F}(r) = \text{P}(X \leq r) = \sum_{i=0}^{r} C_n^i \cdot p^i \cdot q^{n-i}.$$

Binomijakauman todennäköisyysfunktion ja kertymäfunktion arvot on kätevä laskea Excelin funktion BINOMIJAKAUMA (engl. BINOMDIST) avulla. Kun lasketaan todennäköisyysfunktion arvo, funktion muoto on BINOMIJAKAUMA(*r;n;p;*0). Vastaavasti kertymäfunktion arvon laskevan funktion muoto on BINOMIJAKAUMA(*r;n;p;*1). Argumentti r = suotuisten tapahtumien A määrä, n = kaikkien tapahtumien määrä, ja p = suotuisan tapahtuman A todennäköisyys. Kun lasketaan pistetodennäköisyys, viimeisen argumentin arvo on 0 ja, kun lasketaan kertymäfunktion arvo, se on 1.

ESIMERKKI 13

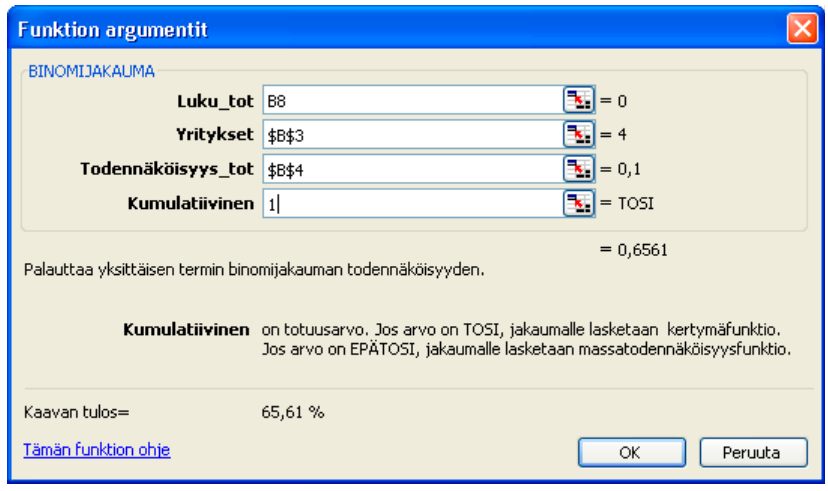
a) Määritä Excelin avulla esimerkin 10 satunnaismuuttujan todennäköisyysjakauma.

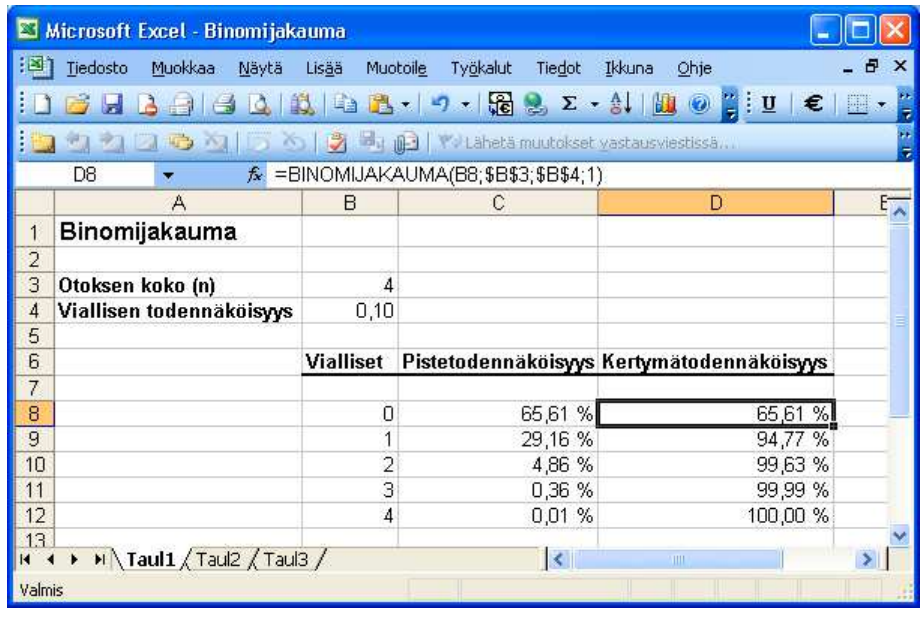
b) Mikä on todennäköisyys, että otoksessa on vähintään 2 viallista?

Satunnaismuuttuja noudattaa jakaumaa Bin(4; 0,1). Lasketaan aluksi Excelin avulla todennäköisyys, että saadaan 0 viallista ja vastaava kertymäfunktion arvo. Sen jälkeen saadaan Kopioi-komennon avulla muut pistetodennäköisyydet ja kertymäfunktion arvot. Funktioiden arvoja laskettaessa kannattaa käyttää soluviittauksia.

(Esimerkki jatkuu seuraavalla sivulla.)

a)





Taulukossa todennäköisyydet on ilmoitettu prosenttilukuina.

b) Tapahtuman P($X \geq 2$) komplementtitapahtuma on, että saadaan enintään yksi viallinen. Sen todennäköisyys on taulukon mukaan 94,77 %, joten kysytty todennäköisyys

$P(X \geq 2) = 100\ \% - 94{,}77\ \% = 5{,}23\ \% = 0{,}0523.$

ESIMERKKI 14

Mikä on esimerkin 10 satunnaismuuttujan X
a) odotusarvo,
b) keskihajonta?

a) Odotusarvon määritelmän mukaan

$$EX = \mu = \sum p_i x_i = 0{,}6561 \cdot 0 + 0{,}2916 \cdot 1 + 0{,}0486 \cdot 2 + 0{,}0036 \cdot 3 + 0{,}0001 \cdot 4 = 0{,}40.$$

Toisaalta sama tulos saadaan muodostamalla tulo np:

$$np = 4 \cdot 0{,}1 = 0{,}40.$$

b) Lasketaan aluksi varianssi:

$$D^2X = \sigma^2 = \sum p_i \cdot (x_i - \mu)^2 = 0{,}6561 \cdot (0 - 0{,}4)^2 + 0{,}2916 \cdot (1 - 0{,}4)^2 + 0{,}0486 \cdot (2 - 0{,}4)^2 + 0{,}0036 \cdot (3 - 0{,}4)^2 + 0{,}0001 \cdot (4 - 0{,}4)^2 = 0{,}36$$

Keskihajonta $DX = \sigma = \sqrt{0{,}36} = 0{,}6$.

Sama tulos saadaan laskemalla tulon npq neliöjuuri ($q = 1 - p$):

$$\sqrt{npq} = \sqrt{4 \cdot 0{,}1 \cdot 0{,}9} = 0{,}6\,.$$

Binomijakauman odotusarvo ja keskihajonta

Oletetaan, että satunnaismuuttuja $X \sim \mathrm{Bin}(n, p)$. Tällöin X:n odotusarvo (keskiarvo) ja keskihajonta voidaan laskea kaavoista

$$EX = \mu = np; \quad DX = \sigma = \sqrt{npq} = \sqrt{np(1-p)}\,.$$

Poisson-jakauma

Binomitodennäköisyyksien laskeminen on teknisesti hankalaa, kun toistojen lukumäärä n on suuri ja tapahtuman todennäköisyys p on pieni. Parametrien n ja p arvoja ei tällaisissa tapauksissa usein tarkkaan tunneta, mutta odotusarvo on tunnettu. Tällöin voidaan käyttää hyväksi seuraavaa lausetta, jonka avulla voidaan laskea Poisson-jakaumaan (Poisson distribution) liittyviä todennäköisyyksiä. Sen todistaminen sivuutetaan.

Binomi- ja Poisson-jakaumien yhteys

Oletetaan, että satunnaismuuttuja $X \sim \text{Bin}(n, p)$, $n \to \infty$ ja $p \to 0$. Oletetaan myös, että tulo np pysyy vakiona. Tällöin

$$\text{P}(X = r) = C_n^r \cdot p^r \cdot q^{n-r} \to \frac{(np)^r}{r!} e^{-np}.$$

Vakio e on Neperin luku ($e \approx 2{,}71828$).

ESIMERKKI 15

Tarkastellaan viittäsataa ydinvoimalaa. Todennäköisyys, että satunnaisesti valitussa ydinvoimalassa sattuu vuoden aikana raportoitava käyttöhäiriö, on 0,01. Mikä on todennäköisyys, että viidessäsadassa ydinvoimalassa sattuu vuoden aikana 7 raportoitavaa käyttöhäiriötä?

Olkoon satunnaismuuttuja X vuoden aikana raportoitavien käyttöhäiriöiden määrä. Tällöin $X \sim \text{Bin}(500; 0{,}01)$. Tulo $np = 500 \cdot 0{,}01 = 5$. Kun käytetään edellä olevaa binomi- ja Poisson-jakaumien yhteyttä saadaan

$$\text{P}(X = 7) = \frac{5^7}{7!} \cdot e^{-5} \approx 0{,}1044 = 10{,}44\ \%.$$

Vertailun vuoksi voidaan todeta, että binomijakauman todennäköisyysfunktiota käyttämällä saadaan tulokseksi 0,1048. Ero on käytännössä merkityksetön.

Poisson-jakauman pistetodennäköisyysfuktio

Olkoon X diskreetti satunnaismuuttuja, joka voi saada arvot 0, 1, 2,..., n. Jos satunnaismuuttujan X todennäköisyysfunktio on

$$P(X = r) = \frac{\delta^r}{r!} \cdot e^{-\delta} \qquad r = 0, 1, 2, 3, \ldots, n,$$

sanotaan, että X noudattaa Poisson-jakaumaa parametrina δ (lue: delta), jota merkitään: $X \sim \text{Poi}(\delta)$. Poisson-jakaumaa noudattavan satunnaismuuttujan odotusarvo

$EX = \delta$ ja keskihajonta $DX = \sqrt{\delta}$.

Käytännössä Poisson-jakaumaa käytetään harvinaisten tapahtumien todennäköisyyksien laskemiseen, kun toistojen määrä n on suuri.

ESIMERKKI 16

Alueella sattuu vuosittain keskimäärin 35 kuolemaan johtavaa liikenneonnettomuutta. Mikä on todennäköisyys, että ensi vuonna kuolemaan johtavien liikenneonnettomuuksien määrä on 25?

Olkoon X kuolemaan johtavien liikenneonnettomuuksien määrä tiettynä vuonna. Koska todennäköisyys joutua tällaiseen onnettomuuteen yhdellä ajokerralla on hyvin pieni ja vuotuisten ajokertojen määrä erittäin suuri, on $X \sim \text{Poi}(\delta)$. Parametrin δ arvo on 35.

$$P(X = 25) = \frac{35^{25}}{25!} \cdot e^{-35} \approx 0,0162 = 1,62\ \%$$

Myös Poisson-jakauman todennäköisyydet on kätevää laskea Excelin avulla.

Poisson-jakauman todennäköisyys lasketaan Excelin funktion POISSON(r;δ;0) (engl. myös POISSON) avulla ja kertymäfunktion arvot saadaan funktion POISSON(r;δ;1) avulla. Funktion argumentti r = suotuisten tapausten lukumäärä ja argumentti δ = Poisson-jakaumaa noudattavan satunnaismuuttujan odotusarvo.

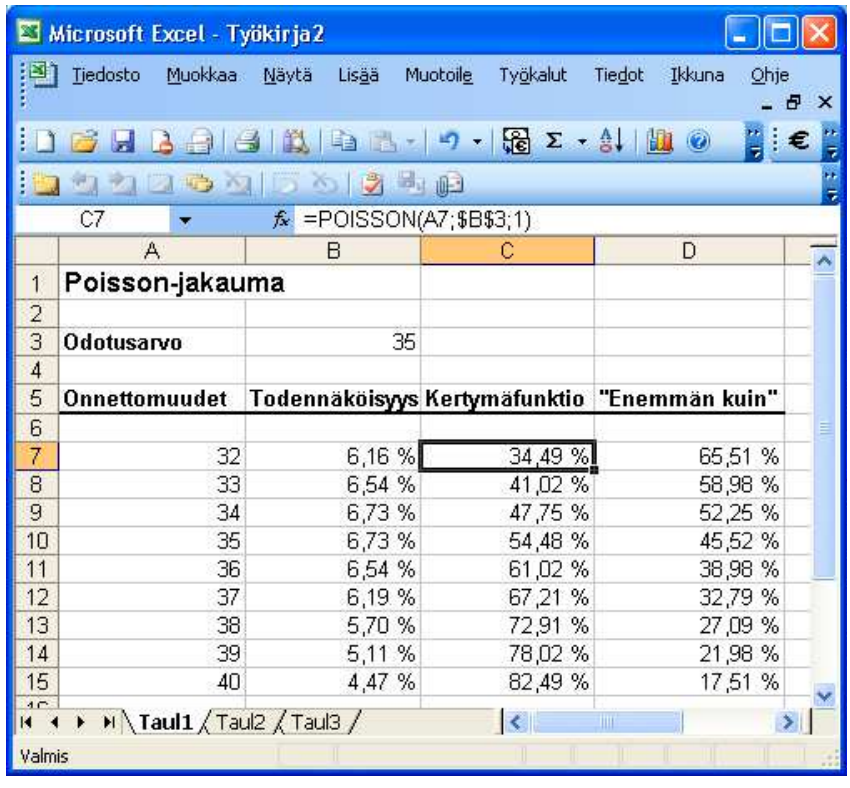
ESIMERKKI 17

a) Laske Excelin avulla edellisen esimerkin todennäköisyys, että alueella tiettynä vuonna kuolemaan johtavien liikenneonnettomuuksien määrä on 32.

b) Mikä on todennäköisyys, että alueella sattuu jonakin yksittäisenä vuonna enintään 32 kuolemaan johtavaa liikenneonnettomuutta?

c) Mikä on todennäköisyys, että jonakin yksittäisenä vuonna sattuu enemmän kuin 40 kuolemaan johtavaa liikenneonnettomuutta?

Tehdään seuraava taulukko. Taulukon ensimmäisen rivin laskemisen jälkeen muut rivit saadaan kopioimalla.



	A	B	C	D
1	**Poisson-jakauma**			
2				
3	**Odotusarvo**	35		
4				
5	**Onnettomuudet**	**Todennäköisyys**	**Kertymäfunktio**	**"Enemmän kuin"**
6				
7	32	6,16 %	34,49 %	65,51 %
8	33	6,54 %	41,02 %	58,98 %
9	34	6,73 %	47,75 %	52,25 %
10	35	6,73 %	54,48 %	45,52 %
11	36	6,54 %	61,02 %	38,98 %
12	37	6,19 %	67,21 %	32,79 %
13	38	5,70 %	72,91 %	27,09 %
14	39	5,11 %	78,02 %	21,98 %
15	40	4,47 %	82,49 %	17,51 %

Taulukossa kaikki todennäköisyydet ovat havainnollisessa prosenttilukumuodossa. Taulukosta voidaan lukea, että esimerkiksi 32 kuolemaan johtavan liikenneonnettomuuden todennäköisyys on 6,16 %. Enintään 32 kuolemaan johtavaa liikenneonnettomuutta sattuu todennäköisyydellä 34,49 % ja enemmän kuin 40 kuolemaan johtavaa liikenneonnettomuutta todennäköisyydellä 17,51 %. Jälkimmäinen luku saadaan komplementtitodennäköisyytenä.

6.2 Jatkuvat todennäköisyysjakaumat

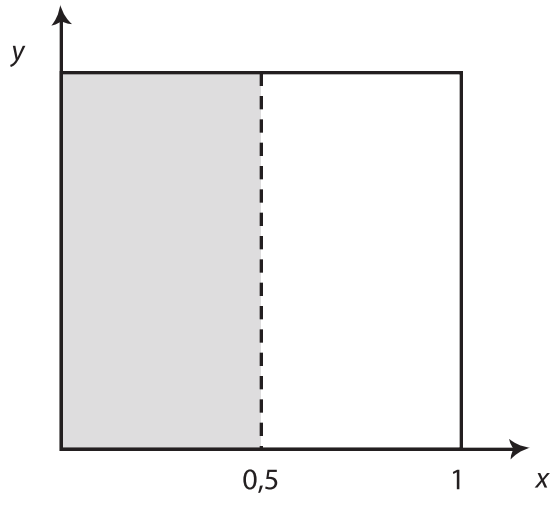
ESIMERKKI 18

Valitaan mielivaltaisesti reaaliluku väliltä [0, 1]. Mikä on todennäköisyys, että saadaan a) luku 0,5, b) enintään 0,5?

a) Olkoon X "valituksi tuleva luku". Koska välillä [0, 1] on äärettömän monta reaalilukua, on $P(X = 0{,}5) = 0$.

b) Tapaus $X \leq 0{,}5$ tarkoittaa, että $0 \leq X \leq 0{,}5$. Koska tämä väli on puolet välistä [0, 1], on $P(X \leq 0{,}5) = 0{,}5$.

Tämä voidaan esittää graafisesti siten, että välin [0, 1] ajatellaan vastaavan aluetta, jonka pinta-ala on yksi pinta-alayksikkö. Kysyttyä todennäköisyyttä vastaa varjostetun alueen pinta-ala.



Edellisen esimerkin satunnaismuuttuja on jatkuva, eli X voi saada minkä tahansa arvon väliltä [0, 1]. Jatkuvalle satunnaismuuttujalle on myös ominaista, että kaikki pistetodennäköisyydet ovat nollia: $P(X = a) = 0$.

Koska jatkuvaan satunnaismuuttujaan liittyvät pistetodennäköisyydet ovat nollia, jatkuvaa todennäköisyysjakaumaa ei voida esittää pistetodennäköisyysfunktion avulla. Sen vastine, kun satunnaismuuttuja on jatkuva, on **tiheysfunktio** (density function). Edellinen kuvio esittää graafisesti esimerkin 17 satunnaismuuttujan tiheysfunktion. Matemaattisesti se voidaan määritellä seuraavasti:

$$f(x) = \begin{cases} 1, & kun\ 0 \leq x \leq 1 \\ 0, & kun\ x < 0\ tai\ x > 1 \end{cases}$$

Tiheysfunktion ominaisuuksia

Olkoon $f(x)$ jatkuvan satunnaismuuttujan X tiheysfunktio. Tällöin

1. $f(x) \geq 0$ kaikilla x:n arvoilla
2. $f(x)$:n kuvaajan ja x-akselin väliin jäävän alueen kokonaispinta-ala on aina yksi.

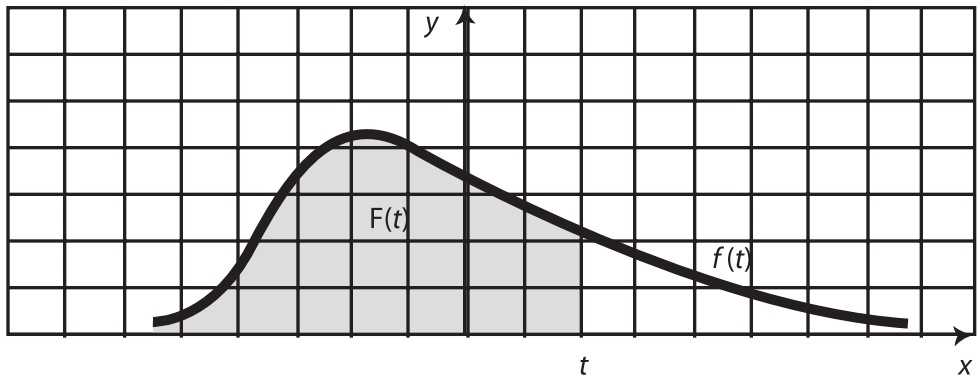
Esimerkin 18 todennäköisyys $P(X \leq 0{,}5)$ saatiin laskemalla sen alueen pinta-ala, jota rajoittavat tiheysfunktion kuvaaja, x- ja y-akseli sekä suora $x = 0{,}5$. Toisaalta aikaisemman määritelmän mukaan $P(X \leq 0{,}5) = F(0{,}5)$, jossa $F(0{,}5)$ on jakauman kertymäfunktion arvo, kun $x = 0{,}5$.

Jatkuvan jakauman kertymäfunktio

Olkoon X jatkuva satunnaismuuttuja. Tällöin

$$P(X \leq t) = F(t) = \int_{-\infty}^{t} f(x)dx\,.$$

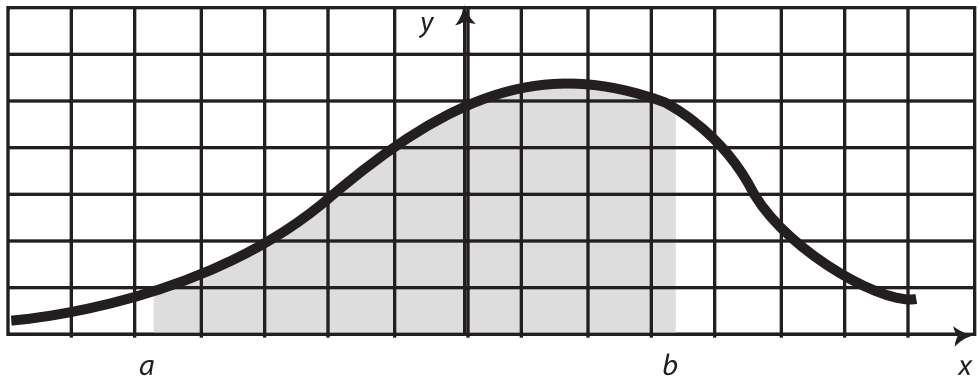
Kuten integraalilaskentaa hallitsevat tietävät, kertymäfunktion arvo on sen alueen pinta-ala, jota rajoittavat tiheysfunktion $f(x)$ kuvaaja, x-akseli ja suora $x = t$ (varjostettu alue).



Jatkuvan jakauman kertymäfunktion määritelmästä seuraa suoraan sen seuraavat ominaisuudet:

Jatkuvan jakauman kertymäfunktion ominaisuuksia

1) $P(a \leq X \leq b) = F(b) - F(a)$
2) $P(X = a) = 0$
3) $P(X \leq a) = P(X < a)$.



$P(a \leq X \leq b) = F(b) - F(a)$

ESIMERKKI 19

Valitaan mielivaltaisesti reaaliluku väliltä [0, 1]. Mikä on todennäköisyys, että saadaan luku, joka on
a) lukujen 0,2 ja 0,5 välillä,
b) vähintään 0,7?

a) Koska $P(X \leq 0{,}2) = F(0{,}2) = 0{,}2$ ja $P(X \leq 0{,}5) = F(0{,}5) = 0{,}5$ (vrt. esimerkkiin 18), on $P(0{,}2 \leq X \leq 0{,}5) = F(0{,}5) - F(0{,}2) = 0{,}3$.

b) $P(X \geq 0{,}7) = 1 - P(X < 0{,}7) = 1 - F(0{,}7) = 1 - 0{,}7 = 0{,}3$.

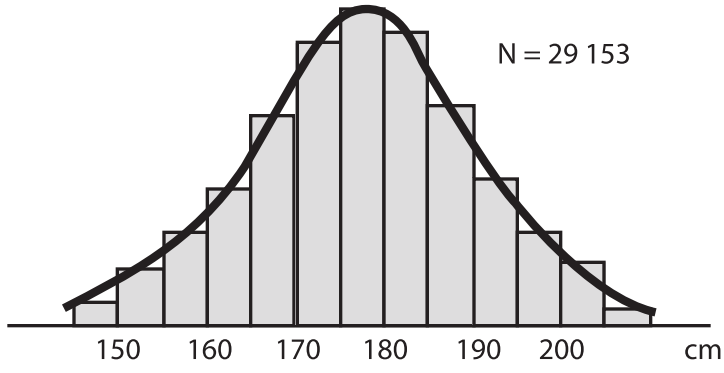
Jatkuvan satunnaismuuttujan odotusarvo ja keskihajonta määritellään vastaavalla tavalla kuin diskreetin jakauman odotusarvo ja keskihajonta korvaamalla summamerkit määrätyillä integraaleilla. Tässä esityksessä määritelmät sivuutetaan.

Normaalijakauma

Tarkastellaan seuraavaksi eräitä aivan keskeisiä jatkuvia jakaumia.

ESIMERKKI 20

Kuviossa esitetään asevelvollisten kutsuntaan eräänä vuonna osallistuneiden miesten pituuksien jakauma. Tämä jakauma on varsin symmetrinen ja valtaosa havaintoarvoista on lähellä keskiarvoa. Kuvioon on piirretty myös **Gaussin käyrä,** joka on lähes yhtenevä empiirisen jakauman kanssa. Jakaumaa, jonka tiheysfunktio on Gaussin käyrä, sanotaan **normaalijakaumaksi** (normal distribution).



Normaalijakauma on yleisin kaikista todennäköisyysjakaumista. Saksalaista matemaatikkoa Karl Friedrich Gaussia (1777–1855) pidetään sen keksijänä. Gauss käytti normaalijakaumaa muun muassa tähtitieteessä mittausvirheiden jakauman mallina.

Myöhemmin on osoitettu, että sellaiset satunnaismuuttujat, joiden arvoon vaikuttavat monet pienet toisistaan riippumattomat tekijät, noudattavat usein normaalijakaumaa (keskeinen raja-arvolause). Käytännössä muun muassa ihmisten henkiset ja fyysiset ominaisuudet, teollisessa tuotannossa syntyvien tuotteiden ominaisuudet ja havaintovirheet missä tahansa mittauksessa ovat likimäärin normaalisti jakautuneita.

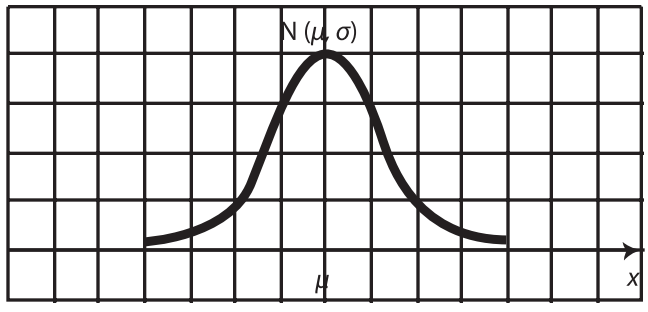
Normaalijakauman määritelmä

Olkoon X jatkuva satunnaismuuttuja, joka voi saada kaikki reaaliarvot välillä $(-\infty, \infty)$. Lisäksi oletetaan, että X:n odotusarvo on μ ja keskihajonta σ. Jos X:n tiheysfunktio on muodoltaan

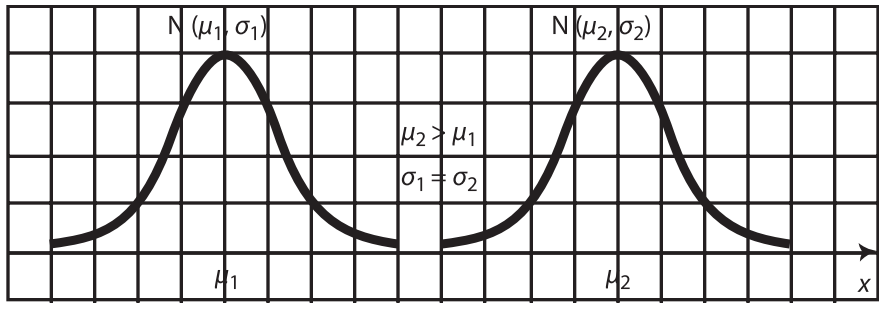
$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma} \right)^2},$$

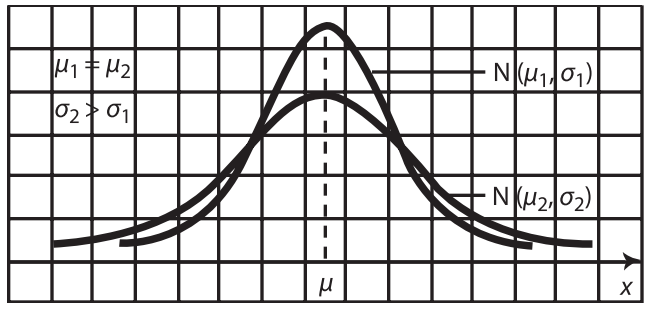
satunnaismuuttuja X noudattaa normaalijakaumaa parametrein μ ja σ. Tätä merkitään: $X \sim N(\mu, \sigma)$. Tiheysfunktion lausekkeessa
μ = odotusarvo (keskiarvo)
σ = (keski)hajonta
π = ympyrän kehän ja halkaisijan suhde ($\pi \approx 3{,}14159$)
e = Neperin luku ($e \approx 2{,}71828$).

Normaalijakauman tiheysfunktion kuvaaja on edellä mainittu Gaussin kellokäyrä.



Normaalijakauman tapauksessa parametrit μ ja σ määrittävät jakauman täysin. Jakauman sijainti x-akselilla riippuu parametrista μ ja jakauman muoto parametrista σ.





Kuten edellisistä kuvioista käy ilmi, normaalijakauma on symmetrinen jakauma. Puolet normaalijakaumaa noudattavan satunnaismuuttujan arvoista on pienempiä kuin odotusarvo ja puolet suurempia kuin odotusarvo.

Normaalijakaumaan liittyvät todennäköisyydet määritetään kertymäfunktion avulla.

Normaalijakauman kertymäfunktio

Normaalijakauman kertymäfunktio on muodoltaan

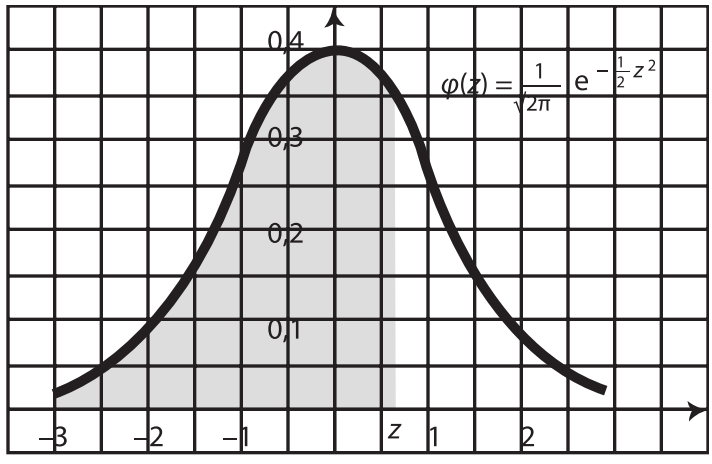
$$\mathrm{P}(X \leq t) = \mathrm{F}(t) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{t} e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma} \right)^2} dx.$$

Koska normaalijakauman tiheysfunktion integroiminen on mahdollista vain numeerisesti, todennäköisyydet lasketaan käytännössä tähän tarkoitukseen laaditun taulukon tai tietokoneen avulla. Taulukoiden tekeminen kaikille parametrien μ ja σ arvoille on kuitenkin mahdotonta. Taulukot onkin laadittu vain sellaiselle satunnaismuuttujalle, joka on jakautunut N(0, 1). Todennäköisyyksien laskeminen perustuu siihen, että mikä tahansa normaalisti jakautunut satunnaismuuttuja voidaan yksinkertaisella muunnoksella **normittaa** siten, että sen jakauma on N(0, 1).

Normitettu satunnaismuuttuja

Olkoon satunnaismuuttuja $X \sim \mathrm{N}(\mu, \sigma)$. Tällöin satunnaismuuttuja $Z = \dfrac{X - \mu}{\sigma}$ on normaalisti jakautunut parametrein 0 ja 1, eli $Z \sim \mathrm{N}(0, 1)$.

Normitetun normaalijakauman tiheysfunktiosta käytetään merkintää $\varphi(z)$ ja kertymäfunktiosta $\Phi(z)$.



Tiheysfunktion $\varphi(z)$ ja z-akselin väliin jäävän alueen kokonaispinta-ala on 1. Varjostettu alue kuviossa vastaa todennäköisyyttä $P(Z \leq z) = \Phi(z)$. Normitetun satunnaismuuttujan Z kertymäfunktion arvot ovat taulukossa kirjan liitteenä (Normitetun normaalijakauman kertymäfunktion arvoja).

ESIMERKKI 21

Olkoon satunnaismuuttuja $Z \sim N(0,1)$. Mikä on todennäköisyys, että

a) $Z < 1{,}2$

b) $Z > 1{,}2$

c) $-1{,}2 < Z < 1{,}2$?

Käyttämällä hyväksi jatkuvan satunnaismuuttujan kertymäfunktion määritelmää ja ominaisuuksia saadaan

a) $P(Z < 1{,}2) = \Phi(1{,}2) = 0{,}8849$ (ks. kirjan liitteenä olevaa taulukkoa)

b) $P(Z > 1{,}2) = 1 - P(Z \leq 1{,}2) = 1 - \Phi(1{,}2) = 1 - 0{,}8849 = 0{,}1151$

c) $P(-1{,}2 \leq Z \leq 1{,}2) = \Phi(1{,}2) - \Phi(-1{,}2) = 0{,}8849 - 0{,}1151 = 0{,}7698.$

ESIMERKKI 22

Tutkimusten mukaan ihmisten älykkyys on jakautunut N(100, 24). Mikä on todennäköisyys, että satunnaisesti valitun henkilön älykkyysosamäärä on

a) enintään 130

b) 76–124?

a) Satunnaismuuttujan arvoa X vastaa normitetun satunnaismuuttujan arvo

$$Z = \frac{130 - 100}{24} = 1{,}25$$

$P(X \leq 130) = P(Z \leq 1{,}25) = \Phi(1{,}25) = 0{,}894$

b) $P(76 \leq X \leq 124) = P\left(\frac{76 - 100}{24} \leq Z \leq \frac{124 - 100}{24}\right) =$

$P(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1) = 0{,}8413 - 0{,}1587 = 0{,}6826$

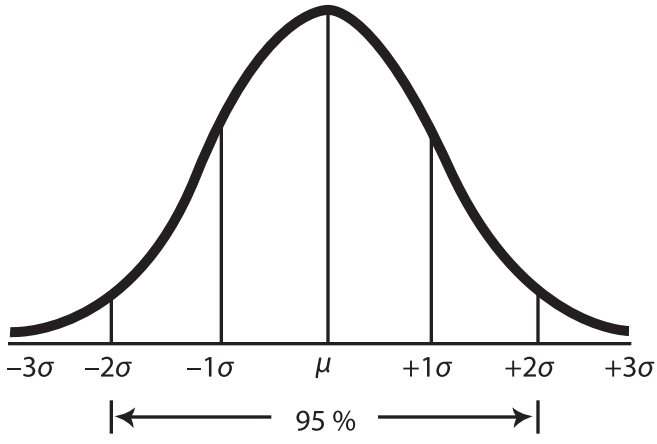
Jälkimmäinen tulos osoittaa, että 68 %:lla ihmisistä älykkyys poikkeaa enintään yhden hajonnan mitan keskiarvosta. Tämä tulos on yleistettävissä.

Normaalijakauman ominaisuuksia

Oletetaan, että satunnaismuuttuja X noudattaa normaalijakaumaa $N(\mu, \sigma)$. Tällöin satunnaismuuttujan arvoista

- 68 % on välillä $[\mu - \sigma, \mu + \sigma]$
- 95 % on välillä $[\mu - 1{,}96\sigma, \mu + 1{,}96\sigma]$
- 99 % on välillä $[\mu - 2{,}58\sigma, \mu + 2{,}58\sigma]$
- 99,9 % on välillä $[\mu - 3{,}30\sigma, \mu + 3{,}30\sigma]$.

Näitä ominaisuuksia käytetään hyväksi tilastollisessa päätöksenteossa.



ESIMERKKI 23

Palataan tämän luvun alussa esitettyyn ongelmaan, jossa oli selvitettävä, kuinka moni 16 vuotta täyttäneestä neljästä miljoonasta suomalaisesta pääsisi älykkyytensä puolesta Mensa ry:n jäseneksi. Vaatimuksena on vähintään pistemäärä 148 älykkyystestissä, jossa kaikkien ihmisten pistemäärien keskiarvo on 100 ja keskihajonta 24.

Kuten edellä todettiin, kansalaisten älykkyys on normaalisti jakautunut. Mensan käyttämässä testissä jakauma on $N(100, 24)$.

Lasketaan aluksi todennäköisyys, että satunnaisesti valittu 16 vuotta täyttänyt henkilö saisi testissä vähintään pistemäärän 148:

$$P(X \leq 148) = 1 - P(X \leq 148) = 1 - P\left(Z < \frac{148 - 100}{24}\right) =$$

$$1 - P(Z < 2) = 1 - \Phi(2) = 1 - 0{,}9772 = 0{,}0228$$

Saatu tulos tarkoittaa, että satunnaisesti valitun henkilön älykkyysosamäärä täyttää vaadittavat ehdot 2,3 %:n todennäköisyydellä. Neljästä miljoonasta 16 vuotta täyttäneestä suomalaisesta siis 2,3 % pääsisi älykkyytensä puolesta Mensa ry:n jäseneksi.

$0{,}023 \cdot 4\,000\,000$ henkilöä $= 92\,000$ henkilöä.

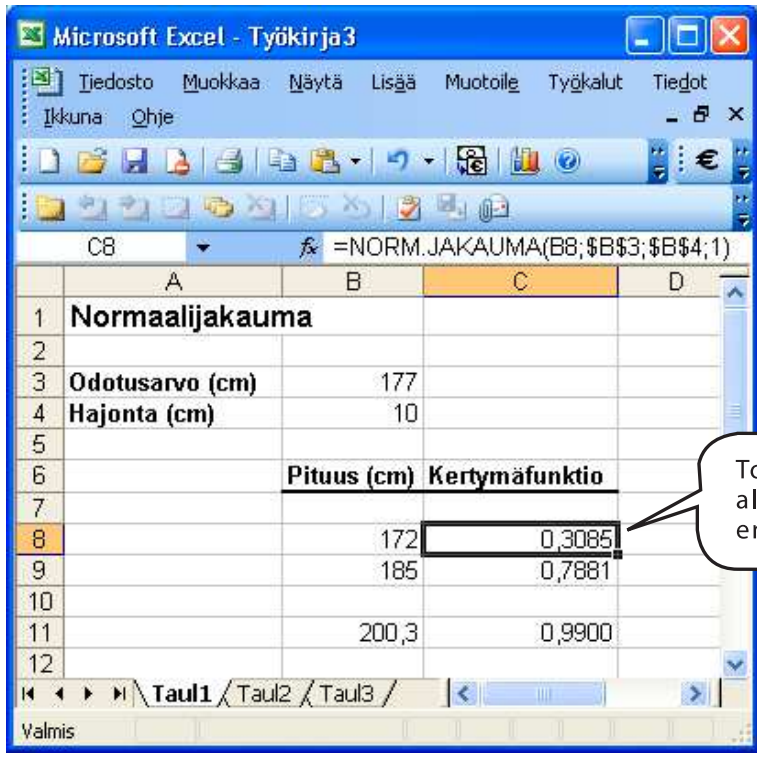
Myös normaalijakaumaan liittyvät todennäköisyydet voidaan laskea Excelin funktioiden avulla. Funktion NORM.JAKAUMA($x;\mu;\sigma;1$) avulla voi laskea normaalijakauman kertymätodennäköisyyksiä. Vastaava funktio Excelin englanninkielisessä versiossa on NORMDIST. Tiettyä todennäköisyyttä vastaavan satunnaismuuttujan arvon voi laskea funktion NORM.JAKAUMA.KÄÄNT($p;\mu;\sigma$), jossa p = kertymätodennäköisyys. Vastaava funktio Excelin englanninkielisessä versiossa on NORMINV.

ESIMERKKI 24

Asepalvelukseen astuvien alokkaiden pituus noudattaa normaalijakaumaa N(177 cm, 10 cm). Mikä on todennäköisyys, että satunnaisesti valitun alokkaan pituus on

a) enintään 172 cm
b) 172 cm – 185 cm
c) yli 185 cm?
d) Mikä on pituus, jonka ylittää 1 % alokkaista?

Kaikki kohdat voidaan laskea seuraavan Excel-taulukon avulla:



a) $P(X \leq 172\ \text{cm}) = F(172) = 0{,}3085 \approx 31\ \%$

b) $P(172\ \text{cm} \leq X \leq 185\ \text{cm}) = F(185) - F(172) = 0{,}7881 - 0{,}3085 = 0{,}4796 \approx 48\ \%$

c) $P(X \geq 185\ \text{cm}) = 1 - P(X < 185\ \text{cm}) = 1 - F(185) = 1 - 0{,}7881 = 0{,}2119 \approx 21\ \%$

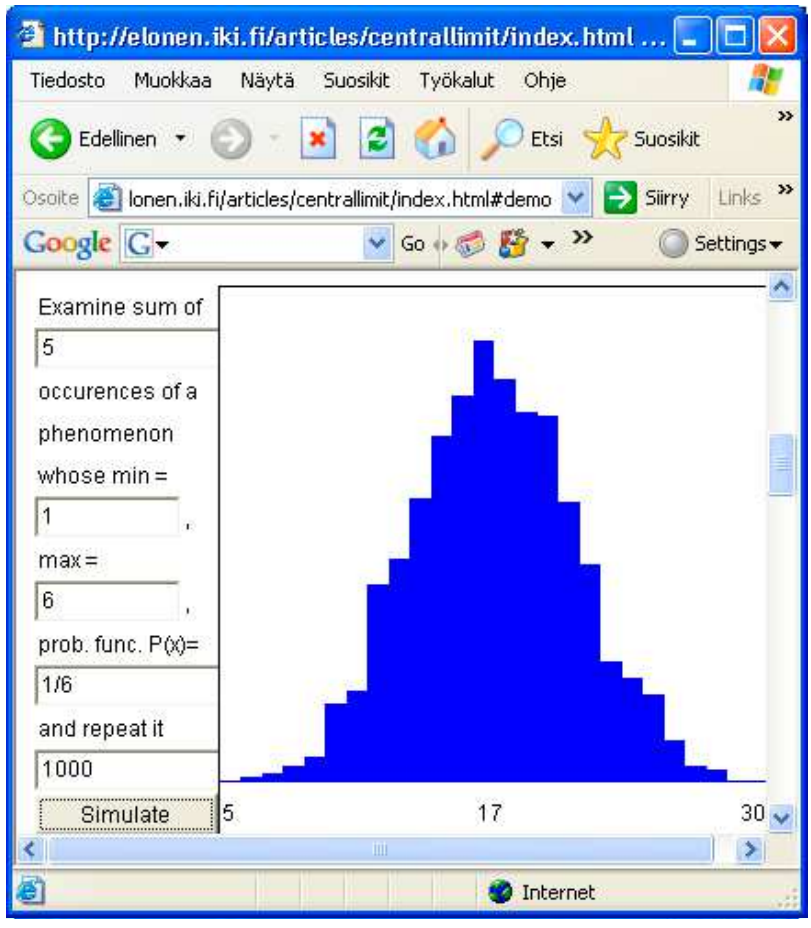
d) Jos alokkaista 1 % on pidempiä kuin tietty pituus a, enintään 99 % heistä on yhtä pitkiä kuin a. Sitä vastaa siis kertymäfunktion arvo 99 % eli desimaalilukuna 0,99. Kertymäfunktion arvoa 0,99 vastaavaksi a:n arvoksi saadaan sijoittamalla soluun B11 kaava = NORM.JAKAUMA.KÄÄNT(C11;B3;B4). Tuloksena on 200,3 cm. Toisin sanottuna 99 % alokkaista on enintään 200 cm pitkiä, joten sitä pidempiä on 1 %.

Yksi tilastotieteen merkittävimmistä lauseista on **keskeinen raja-arvo-lause** (central limit theorem). Muiden muassa monet tilastolliset testit perustuvat tähän lauseeseen.

ESIMERKKI 25

Jos heitetään yhtä noppaa 600 kertaa, jokainen silmäluku esiintyy noin 100 kertaa, jos noppa on virheetön. Nopan heiton silmälukujen jakauma on tasainen. Tilanne muuttuu, jos heitetään esimerkiksi 5 noppaa riittävän monta kertaa ja kirjataan ylös silmälukujen summa. Viiden nopan heitossa silmälukujen summa voi vaihdella lukujen 5 (kaikki silmäluvut ykkösiä) ja 30 (kaikki kuutosia) välillä. Kuitenkin silmälukujen summa 5 on paljon harvinaisempi kuin esimerkiksi silmälukujen summa 17. Silmälukujen summaksi voi tulla 17 monella tavalla ($1 + 4 + 2 + 6 + 4 = 17$ tai $3 + 3 + 3 + 6 + 2 = 17$ jne.), mutta silmälukujen summa 5 saadaan vain, jos kaikkien noppien silmäluku on 1.

Internetissä on saatavilla useita simulointiohjelmia, joilla voi kokeilla muun muassa nopanheiton simulointia. Seuraava tulos on saatu ”heittämällä” viittä noppaa 1 000 kertaa ja kirjaamalla kustakin heitosta noppien silmälukujen summa.



Silmälukujen summan jakauma muistuttaa normaalijakaumaa, mikä ei ole suinkaan sattuma. Jos heittojen lukumäärää edelleen kasvatetaan, empiirinen jakauma lähenee yhä enemmän normaalijakauman muotoa.

Keskeinen raja-arvolause

Oletetaan, että satunnaismuuttujat $X_1, X_2, \ldots X_n$ ovat riippumattomia ja noudattavat kukin samaa mielivaltaista jakaumaa. Tällöin summamuuttuja $X = X_1 + X_2 + \ldots + X_n$, joka on myös satunnaismuuttuja, noudattaa likimain normaalijakaumaa, kun n on riittävän suuri. Satunnaismuuttujan X jakauma on sitä tarkemmin normaalinen, mitä suurempi n on.

ESIMERKKI 26

Sarjatuotannossa syntyy keskimäärin 10 % viallisia tuotteita. Poimitaan laaduntarkkailua varten tuotannosta 50 kappaletta toisistaan riippumattomia kahden kappaleen otoksia. Merkitään viallisten tuotteiden määrää kussakin otoksessa $X_1, X_2, \ldots, X_{50}$. Mikä on satunnaismuuttujan $X = X_1 + X_2 + \ldots + X_{50}$ jakauma?

Jokainen satunnaismuuttujista $X_1, X_2, \ldots, X_{50}$ on jakautunut Bin(2; 0,1). Toisaalta keskeisen raja-arvolauseen mukaan satunnaismuuttujien summa X noudattaa jakaumaa N(μ, σ). Määritetään μ ja σ:

$EX = EX_1 + EX_2 + \ldots + EX_{50} = 2 \cdot 0{,}1 + 2 \cdot 0{,}1 + \ldots + 2 \cdot 0{,}1 = 50 \cdot 2 \cdot 0{,}1 = 10$. Siis $\mu = 10$.

$D^2X = D^2X_1 + D^2X_2 + \ldots + D^2X_{50} = 2 \cdot 0{,}1 \cdot 0{,}9 + 2 \cdot 0{,}1 \cdot 0{,}9 + \ldots + 2 \cdot 0{,}1 \cdot 0{,}9 = 50 \cdot 2 \cdot 0{,}1 \cdot 0{,}9 = 9$.

Siis $\sigma^2 = 9$ eli $\sigma = 3$.

Satunnaismuuttuja $X = X_1 + X_2 + \ldots + X_{50}$ on näin ollen jakautunut likimain N(10, 3). Toisaalta esimerkistä voidaan päätellä, että X noudattaa jakaumaa Bin(100; 0,1). Tämä tulos on yleistettävissä.

Binomi- ja normaalijakauman yhteys

Oletetaan, että satunnaismuuttuja $X \sim \text{Bin}(n, p)$. Jos n on riittävän suuri, X noudattaa likimäärin jakaumaa

$\text{N}(np, \sqrt{npq})$, jossa $q = 1 - p$.

On mahdotonta sanoa tarkasti, mistä n:n arvosta lähtien edellinen tulos on voimassa, koska voimassaolo riippuu myös todennäköisyydestä p. Jos $p \approx 0{,}5$, tulos on riittävän tarkasti voimassa, kun $n \approx 30$.

Jos $p \approx 0{,}1$, n:n on oltava vähintään 100. Yleisesti binomijakauman ja normaalijakauman yhteys on riittävän tarkka, kun ehto $n \cdot p \cdot (1 - p) \approx 9$ on voimassa.

ESIMERKKI 27

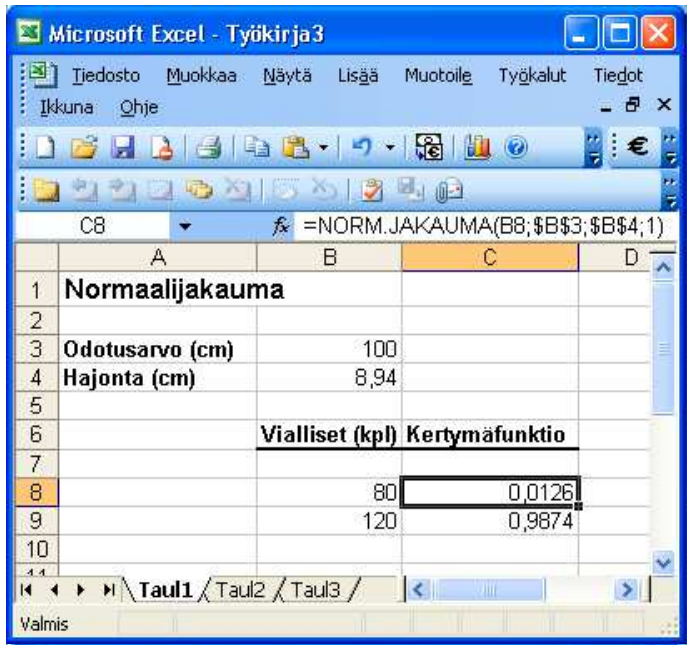
Tuotteista on viallisia 20 %. Otetaan 500 kappaleen satunnaisotos. Mikä on todennäköisyys, että viallisia on a) enintään 80, b) 80–120?

Olkoon X viallisten tuotteiden lukumäärä. Tällöin $X \sim \text{Bin}(500; 0{,}2)$. Tehtävän ratkaiseminen binomijakauman todennäköisyysfunktion avulla on kuitenkin hyvin työlästä.

Edellisen lauseen mukaan X noudattaa likimäärin jakaumaa

$\text{N}(500 \cdot 0{,}2; \sqrt{500 \cdot 0{,}2 \cdot 0{,}8}) = \text{N}(100; 8{,}94)$.

Kysytyt todennäköisyydet voidaan laskea normaalijakauman avulla:



a) $P(X \leq 80 \text{ kpl}) = F(80) = 0{,}0126 \approx 1{,}3\ \%$.

b) $P(80 \leq X \leq 120) = F(120) - F(80) = 0{,}9874 - 0{,}0126 = 0{,}9748 \approx 97{,}5\ \%$.

ESIMERKKI 28

Tutkimuksessa pyritään selvittämään opiskelijoiden älykkyyden jakaumaa. Etukäteen tiedetään, että älykkyysosamäärä on normaalisti jakautunut, mutta parametrien μ ja σ arvoa ei tunneta. Parametrien arvot pyritään määrittämään poimimalla perusjoukosta viisi kappaletta riippumattomia otoksia. Kuhunkin otokseen kuuluvien opiskelijoiden älykkyysosamäärä tutkitaan, minkä jälkeen lasketaan jokaisesta otoksesta älykkyysosamäärien keskiarvo. Keskiarvot olivat 112,4; 110,1; 105,6; 109,6 ja 107,5. Myös otoskeskiarvo on satunnaismuuttuja, jonka arvo vaihtelee eri otoksissa. Sen jakaumaa sanotaan keskiarvon $\overline{x}$ **otantajakaumaksi**.

Keskiarvon otantajakauma

Oletetaan, että satunnaismuuttuja X on jakautunut $N(\mu, \sigma)$. Poimitaan jakaumasta n kappaleen satunnaisotos. Tällöin voidaan osoittaa, että otoskeskiarvo $\overline{x} = \frac{\sum x_i}{n}$ noudattaa jakaumaa $N(\mu, \frac{\sigma}{\sqrt{n}})$ ja normitettu satunnaismuuttuja $\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$.

Otoskeskiarvojen keskihajontaa $\frac{\sigma}{\sqrt{n}}$ sanotaan keskiarvon **keskivirheeksi**.

Käytännössä perusjoukon keskihajontaa ei yleensä tunneta. Tällöin se korvataan otoksesta lasketulla keskihajonnalla s. Jos n on pieni ($n < 30$), keskiarvon jakauma ei kuitenkaan ole välttämättä normaalinen.

Studentin t-jakauma

Oletetaan, että satunnaismuuttuja X on jakautunut $N(\mu, \sigma)$. Poimitaan jakaumasta n kappaleen satunnaisotos. Tällöin satunnaismuuttuja

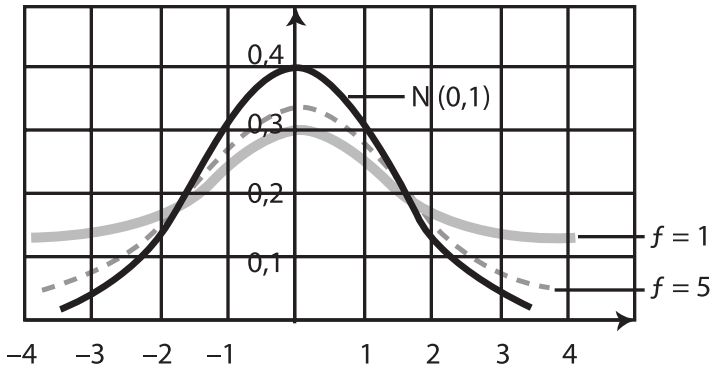
$$t = \frac{x - \mu}{\frac{s}{\sqrt{n}}}$$

noudattaa Studentin t-jakaumaa (Student's distribution) **vapausastein $f = n-1$**.

Tilasto-ohjelmissa vapausasteita merkitään yleensä *df*:llä.

Kaavassa $s = \sum_{i=1}^{n} \frac{\sqrt{\Sigma(x_i - \bar{x})^2}}{n-1}$ on otoksesta laskettu keskihajonta.

Vapausasteet määräävät t-jakauman muodon. Tämä jakauma, kuten normaalijakaumakin, on symmetrinen. Seuraavaan kuvioon on piirretty normaalijakauman ja t-jakauman tiheysfunktioiden kuvaajat vapausastein $f = 5$ ja $f = 1$.



Studentin t-jakaumaa käytetään tilastollisessa päätöksenteossa, kun otoskoko on pieni. Sitä varten t-jakaumalle on laadittu taulukoita (ks. kirjan lopusta liitettä Studentin t-jakaumaan liittyvät kriittiset arvot).

Studentin t-jakauman tiheysfunktion kuvaaja muistuttaa normaalijakauman N(0, 1) tiheysfunktiota sitä enemmän, mitä suurempi vapausasteiden lukumäärä f on. Kun vapausasteiden lukumäärä on vähintään 30, t-jakaumaan liittyvät todennäköisyydet voidaan määrittää jakauman N(0, 1) avulla.

χ^2-jakauma

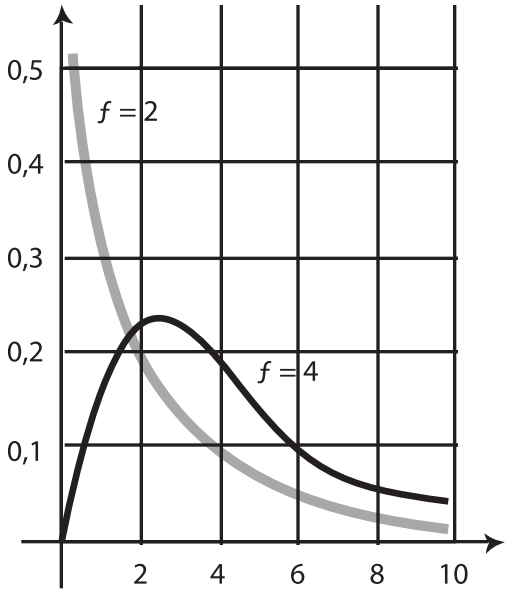
Myös χ^2-jakauma (lue: khi toiseen -jakauma; khi distribution) voidaan johtaa normaalijakaumasta.

Oletetaan, että satunnaismuuttujat $X_1, X_2, ..., X_k$ ovat riippumattomia ja niiden kaikkien jakauma on N(0,1). Tällöin satunnaismuuttuja

$\chi^2 = X_1^2 + X_2^2 + ... + X_k^2$ noudattaa χ^2-jakaumaa vapausastein $f = k$.

χ^2-jakaumaa käytetään muun muassa silloin, kun testataan ristiintaulukoitujen muuttujien riippuvuutta.

χ^2-jakauman muoto vaihtelee huomattavasti vapausasteiden mukaan. Seuraavaan kuvioon on piirretty tämän jakauman tiheysfunktioiden kuvaajat, kun f saa arvot 2 ja 4.



TEHTÄVIÄ

1. Oletetaan, että perheessä on kolme lasta. Muodosta satunnaismuuttujan X = poikien lukumäärä todennäköisyysjakauma ja jakauman kertymäfunktio. Piirrä kertymäfunktion kuvaaja.

2. Mikä on todennäköisyys, että kolmelapsisessa perheessä on
a) 1 tai 2 poikaa,
b) vähintään 2 tyttöä?

3. Erään laboratoriokokeen onnistumistodennäköisyys on 0,7. Koe tehdään kolme kertaa peräkkäin. Olkoon satunnaismuuttuja X kokeen onnistumiskertojen lukumäärä. Määritä satunnaismuuttujan todennäköisyysjakauma ja kertymäfunktio.

4. Mikä on todennäköisyys, että edellisen tehtävän koe
a) onnistuu 2 tai 3 kertaa,
b) onnistuu vähintään kerran,
c) epäonnistuu vähintään kerran?

5. Heitetään yhtä aikaa kahta noppaa. Merkitään X:llä silmälukujen summaa. Laske todennäköisyydet
a) $P(X \leq 4)$,
b) $P(3 < X \leq 10)$,
c) $P(X > 8)$.

6. Hyväntekeväisyysjärjestön arpajaisissa on 1 000 arpaa. Voittoarpoja on yhteensä 31, joista 100 euron voittoja on yksi, 50 euron voittoja 10 ja 20 euron voittoja 20. Määritä arvan hinta siten, että voiton odotusarvo on 55 % arvan hinnasta, kun ostetaan yksi arpa.

7. Maahantuojan pitämän tilaston mukaan erään automallin takuuaikana 50 prosenttiin autoista ei tarvitse tehdä lainkaan takuun piiriin kuuluvia korjauksia. 20 prosenttiin autoista joudutaan tekemään keskimäärin 150 euroa maksavat korjaukset, 25 prosenttiin keskimäärin 250 euroa maksavat korjaukset ja loppuihin 800 euroa maksavat korjaukset. Takuusta aiheutuvat kustannukset otetaan huomioon auton hinnoittelussa. Kuinka paljon takuu lisää automallin hintaa?

8. Vakuutusyhtiö myöntää urheilukilpailujen järjestäjälle niin sanotun sadevakuutuksen. Maksettava korvaus riippuu kilpailun loppumista edeltävän 12 tunnin sademäärästä seuraavasti:

Sademäärä (mm)	Korvaus (€)
0–2	–
3–5	100
6–10	500
11–	1 000

Säätilastojen mukaan 0–2 mm:n sademäärän todennäköisyys on 50 %. Vastaavasti 3–5 mm:n sademäärän todennäköisyys on 30 %, 6–10 mm:n 18 % ja vähintään 11 mm:n 2 %. Määritä vakuutusmaksu siten, että se on 60 % suurempi kuin vakuutuksen perusteella maksettavan korvauksen odotusarvo.

9. Olkoon X ~ Bin(6; 0,4). Muodosta satunnaismuuttujan todennäköisyysjakauma ja esitä se graafisesti.

10. Tilastojen mukaan erään yrityksen tilivuosista 15 % on 80 vuoden aikana ollut tappiollisia. Mikä on todennäköisyys, että 10 vuodesta
a) 2 on tappiollista,
b) enintään 2 on tappiollista?

11. Mikä on edellisessä tehtävässä tappiollisten tilivuosien lukumäärän odotusarvo?

12. Yritys aloittaa uuden kestokulutustavaran maahantuonnin. Tuotteen ulkomainen valmistaja ilmoittaa, että tilastojen mukaan takuun piiriin kuuluvia vikoja on ollut 5 prosentissa tuotteista. Yritys haluaa mitoittaa takuun perusteella tehtävät korjaukset siten, että korjaamon kapasiteetti on 20 % suurempi kuin odotettavissa oleva viallisten määrä. Kuinka suuri korjaamon kapasiteetin pitäisi olla, kun tilivuoden myynniksi on budjetoitu 10 000 kappaletta?

13. Vuoden kuluttua ulkomainen valmistaja ilmoittaa edellisen tehtävän yritykselle, että tehtaan laaduntarkkailun paranemisen vuoksi viallisten tuotteiden määrä on vähentynyt neljään prosenttiin. Takuukorjausten kapasiteetti päätetään mitoittaa nyt niin, että se on takuuaikana odotettavissa oleva viallisten määrä + kaksi kertaa keskihajonta. Kuinka suuri korjaamon kapasiteetin pitäisi nyt olla, kun tilivuoden myynnin arvioidaan kasvavan 20 % edellisestä tilivuodesta?

14. Veroviraston asiakkaista keskimäärin yksi viidestä on ”hankala”. Erään tunnin aikana kävi 8 asiakasta. Mikä on todennäköisyys, että
a) kaksi asiakkaista oli hankalia,
b) enintään kaksi asiakkaista oli hankalia,
c) vähintään kaksi asiakkaista oli hankalia?

15. Postimyyntiliikkeen lähettämistä tavaroista keskimäärin 15 kappaletta kuukaudessa ei tavoita vastaanottajaa. Mikä on todennäköisyys, että ensi kuussa
a) 10 lähetystä,
b) 20 lähetystä ei tavoita vastaanottajaa?

16. Harvinaiseen tautiin sairastuu vuosittain keskimäärin 10 ihmistä. Mikä on todennäköisyys, että seuraavana vuonna sairastuneita on
a) 8,
b) enintään 4,
c) vähintään 3?

17. Televisiomainoksen nähneistä 4 % ostaa tuotteen. Työpaikan 40 ihmisestä 60 % näki mainoksen. Mikä on todennäköisyys, että heistä
a) yksi,
b) enintään yksi,
c) vähintään yksi ostaa tuotteen?

18. Yleiseen hälytysnumeroon tulevista soitoista aiheettomia on keskimäärin 0,15 %. Kuukauden aikana hälytysnumeroon soitetaan keskimäärin 1 200 puhelua. Mikä on todennäköisyys, että seuraavassa kuussa aiheettomien soittojen määrä on
a) 2,
b) enintään 3,
c) vähintään 5?

19. Puhelinkeskukseen tulevien puhelujen määrä noudattaa ns. Poissonin jakaumaa: todennäköisyys, että minuutissa tulee n (≥ 0) puhelua, on $p_n = \frac{a^n}{n!}e^{-a}$, jossa vakio a kuvaa keskuksen ruuhkaisuutta. Laske todennäköisyys, että keskukseen tulee ainakin 5 puhelua, kun $a = 3$. (Yo 2000 K, pitkä)

20. Uusi hampurilaisravintola mitoitettiin palvelemaan neljää asiakasta minuutissa. Ensimmäisinä päivinä ravintolassa kävi lounasaikaan keskimäärin kolme asiakasta minuutissa. Arvioi Poissonin jakaumaa käyttäen todennäköisyys, että ravintolan palvelukapasiteetti lounasaikaan ylittyy. (Yo 2004 S, lyhyt)

21. Oletetaan, että satunnaismuuttuja X on jakautunut N(20, 3) ja satunnaismuuttuja Y ~ N(15, 6). Vertaile satunnaismuuttujien tiheysfunktioiden kuvaajia.

22. Vertailutestissä todettiin sähkölamppumerkin A kestoiän (h) jakaumaksi N(2300,150). Vastaavasti merkin B kestoiän jakauma on N(1 800, 50). Arvioi kumpia lamppuja kannattaa ostaa, jos halutaan saada
a) mahdollisimman tasalaatuisia tuotteita,
b) mahdollisimman kauan kestävä lamppu?

Määritä seuraavat tehtävät kirjan liitteenä olevan taulukon tai Excelin funktioiden avulla:

23. Olkoon $Z \sim N(0, 1)$. Laske
a) $P(Z \leq 0{,}82)$,
b) $P(Z \geq 0{,}82)$,
c) $P(-1{,}96 \leq Z \leq 1{,}96)$.

24. Olkoon $Z \sim N(0, 1)$. Laske z, kun
a) $P(Z \leq z) = 0{,}43$,
b) $P(Z \geq z) = 0{,}43$,
c) $P(-z \leq Z \leq z) = 0{,}95$.

25. Olkoon $X \sim N(117, 23)$. Laske
a) $P(X \leq 123)$,
b) $P(X \geq 110)$,
c) $P(100 \leq X \leq 140)$.

26. Olkoon $X \sim N(\mu, 10)$. Määritä μ siten, että $P(X \leq 80) = 0{,}85$.

27. Olkoon $X \sim N(8, \sigma)$. Määritä σ siten, että $P(X \leq 10) = 0{,}70$.

28. Määritä edellisen tehtävän jakaumasta
a) $P(X \leq 6{,}5)$,
b) $P(6{,}5 \leq X \leq 11{,}7)$.

29. Sähkölamppujen kestoiän jakauma on likimäärin normaalinen parametrein $\mu = 2\,000$ h, $\sigma = 250$ h. Mikä on todennäköisyys, että mielivaltaisesti valittu lamppu palaa
a) vähintään 1 800 tuntia,
b) 1 800–2 200 tuntia?

30. Rengasliikkeen varastossa on 280 kappaletta henkilöauton renkaita. Tämän tyyppisten renkaiden kestävyys noudattaa likimäärin normaalijakaumaa parametrein $\mu = 30\,000$ km, $\sigma = 5\,000$ km.

a) Määritä todennäköisyys, että satunnaisesti valitut neljä rengasta kestävät kaikki vähintään 28 000 kilometrin ajomatkan.

b) Kuinka monelta tämän erän renkaalta voidaan odottaa vähintään 36 000 kilometrin kestoa?

31. On todettu, että erään televisiomerkin se kestoikä, jona aikana siihen ei tarvitse tehdä mitään korjauksia, on jakautunut N(4,6; 1,2). Parametrit on ilmoitettu vuosina. Kuinka pitkä takuu tälle televisiolle voidaan myöntää, jotta takuukorjauksia joudutaan tekemään enintään 5 %:lle ostetuista televisioista?

32. Tukkuliike ostaa valmistajalta tuotetta 800 kappaleen erinä. Viallisten tuotteiden osuus on keskimäärin 5 %. Laske normaalijakaumaa käyttäen todennäköisyys, että seuraavassa tuote-erässä on enintään 4 % viallisia tuotteita.

33. Auton moottorin sytytystulppia valmistava yritys on testissään todennut, että tulppien kestävyys (km) noudattaa jakaumaa N(45 000, 1 600). Kuinka monen ajokilometrin jälkeen tulpat on vaihdettava, jos halutaan 99,9 %:n varmuus, etteivät tulpat tule käyttökelvottomiksi ennen vaihtoa?

34. Toisen sytytystulppamerkin kestävyys (km) noudattaa jakaumaa N(50 000, 2 000). Autoilija aikoo vaihtaa sytytystulpat huolloissa 40 000 kilometrin välein. Kannattaako hänen ostaa näitä vai edellisessä tehtävässä mainittuja sytytystulppia, jos hän haluaa mahdollisimman suuren varmuuden, etteivät tulpat tule käyttökelvottomiksi ennen vaihtoa?

35. Ammattijärjestön 5 000 jäsentä on ottanut ryhmähenkivakuutuksen. Kuolleisuustilaston perusteella on arvioitu, että ammattijärjestön jäsenellä on 0,12 %:n riski kuolla kalenterivuoden aikana. Millä todennäköisyydellä vakuutusyhtiö joutuu maksamaan korvausta vähintään kymmenestä kuolemantapauksesta vuoden aikana?

36. Psykologinen tutkimuslaitos tekee yritysten toimeksiannosta soveltuvuustestauksia. Pitkäaikaisten tilastojen perusteella voidaan arvioida, että loogista ajattelua mittaavassa testissä on testipistemäärien jakauma N(50, 12). Tietokonealan yrityksen palvelukseen pyrkii 10 henkilöä, joille järjestetään edellä mainittu testi. Yritys ottaa palvelukseen henkilön, jonka testipistemäärä on vähintään 65. Mikä on todennäköisyys, ettei yksikään testatuista 10 henkilöstä täytä yrityksen vaatimuksia?

37. Kaupungin vedenpuhdistamossa veteen lisätään erästä kemikaalia 15 g puhdistettavaan vesimäärään makuvirheiden poistamiseksi. Kemikaalin lisäämisen hoitaa annostelija automaattisesti. Yhdellä kertaa puhdistetaan yhden päivän veden tarve. Tutkimuksessa on todettu, että annostelijan lisäämän kemikaalin määrä noudattaa jakaumaa N(15; 1,2). Parametrit on annettu grammoina. Vedessä esiintyy makuhaittoja, jos annostelijan lisäämän kemikaalin määrä on pienempi kuin 12 g tai suurempi kuin 18 g. Kuinka monena päivänä vuodessa makuvirheitä voidaan olettaa esiintyvän?

38. Suomalaisista koululaisista on erään tilaston mukaan 5 % vasenkätisiä. Laske sen tapahtuman todennäköisyys, että 32 oppilaan luokassa on 4 vasenkätistä. Opastus: Käytä joko binomijakaumaa ($p = 0{,}05$, n = 32) tai normaalijakaumaa, missä $\mu = np$ ja $\sigma^2 = np(1 - p)$ tapahtumana $3{,}5 < X \leq 4{,}5$. (Yo 1995 K, lyhyt)

39. Erään 16 500 ihmistä käsittävän ryhmän pituuden oletetaan noudattavan normaalijakaumaa N(p, σ), jossa $p = 175$ cm ja $\sigma = 5$ cm. Ryhmän älykkyysosamäärän oletetaan samoin jakautuvan normaalisti N(102; 12). Mikä on todennäköisyys, että ryhmästä löytyy ainakin yksi henkilö, jonka pituus on yli 190 cm ja älykkyysosamäärä yli 140? Ryhmän kengännumerot ovat nekin jakautuneet normaalisti N(41,6; 0,6). Voidaanko samalla päättelyllä kuin yllä laskea todennäköisyys, että ryhmässä on ainakin yksi henkilö, jonka pituus on yli 190 cm ja kengännumero vähintään 45? Perustele. (Yo 1997 S, lyhyt)

40. Työhönottohaastattelu kestää keskimäärin 25 minuuttia. Oletetaan, että sen pituus on normaalisti jakautunut N(25, σ). Mikä on keskihajonnan σ oltava, että haastattelu ei 95 prosentin varmuudella ylitä 30 minuuttia? (Yo 1999 S, lyhyt)

41. Todennäköisyys, että asiakas palauttaa postimyyntiliikkeestä tilaamansa tuotteen, on 11 %. Mikä on todennäköisyys, että postimyyntiliikkeen 2 500 asiakkaasta vähintään 300 palauttaa tilaamansa tuotteen? Opastus: Approksimoi binomijakaumaa normaalijakaumalla, jonka tunnuslukuina ovat binomijakauman keskiarvo ja hajonta. (Yo 2003 S, lyhyt)

42. Kilon sokeripusseja täytetään koneellisesti. Pussituskone voidaan asettaa täyttämään pusseja, joiden paino noudattaa normaalijakaumaa keskiarvona koneelle annettu paino ja keskihajontana 8,0 g. Kuinka suureksi sokeripussin paino tulee asettaa koneeseen, jotta 95 % sokeripusseista olisi vähintään 1 000 g? (Yo 2005 K, lyhyt).

43. Henkilö lähtee töihin joka aamu autolla samaan aikaan. Hän saapuu työpaikkansa pysäköintialueelle ajankohtana, joka noudattaa normaalijakaumaa. Keskiarvo on klo 8.50 ja hajonta 5 min. Pysäköintialueelta löytyy paikka 65 %:n todennäköisyydellä, ja sieltä on viiden minuutin kävelymatka työpaikalle. Jos kaikki pysäköintipaikat ovat varattuja, henkilö voi ajaa viidessä minuutissa toiselle alueelle, jolta aina löytyy pysäköintipaikka, mutta jolta on 10 minuutin kävelymatka työpaikalle. Mikä on todennäköisyys, että henkilö saapuu työpaikalleen klo 9.00 jälkeen? (Yo 1996 S, pitkä)

7
Tilastollinen päätöksenteko

Media julkaisee säännöllisesti muun muassa puolueiden kannatuksesta tehtyjä tutkimuksia. Näissä tutkimuksissa kysytään tavallisesti 500–2 000 satunnaisesti valitulta äänioikeutetulta suomalaiselta, mitä puoluetta he äänestäisivät, jos vaalit pidettäisiin nyt. Saadut tulokset yleistetään koskemaan kaikkia äänioikeutettuja.

Eräässä tällaisessa tutkimuksessa suurimman kannatuksen sai puolue A, jota kannatti 24,6 % tutkimuksessa haastatelluista 1 016 äänioikeutetusta. Toiseksi eniten kannatusta sai puolue B, jonka osuus oli 23,0 %. Voidaanko tämän tutkimuksen perusteella suurella varmuudella päätellä, että puolueen A kannatus kaikkien äänioikeutettujen joukossa on suurempi kuin puolueen B kannatus?

Ensimmäinen luotettavien johtopäätösten edellytys on, että tutkimus on tehty **tieteelliselle tutkimukselle asetettujen kriteerien mukaan**. Lisäksi voidaan ajatella, että tuloksen tarkkuus riippuu tiettyyn rajaan saakka **otoksen koosta**.

Jos haastatellaan kaikkia äänioikeutettuja ja he vastaavat rehellisesti, tulos on tietenkin täysin luotettava. Kun tutkimukseen osallistui 1 016 äänioikeutettua, puolueiden todellisen ja tutkimuksessa saadun kannatuksen välillä voi olla ainakin muutaman prosenttiyksikön ero. On nimittäin helppo kuvitella, että jos samanaikaisesti poimittaisiin satunnaisesti toinen yhtä iso äänioikeutettujen joukko ja kysyttäisiin heidän mielipidettään, tulokset eivät olisi ainakaan desimaalien tarkkuudella samoja.

Yksi tapa ratkaista ongelma onkin se, että otettaisiin yhtä aikaa esimerkiksi 5 toisistaan riippumatonta otosta, jotka kaikki tutkittaisiin erikseen. Jos kaikki tutkimukset antaisivat tuloksen, että puolueen A kannatus on suurempi kuin puolueen B kannatus, voitaisiin olla melko varmoja, että tulos pätee myös kaikkien äänestäjien joukossa.

Edellä mainittu tapa tutkia asian todenperäisyys on kuitenkin kallis ja monesti aikaa vievä. Käyttämällä tilastollisen päätöksenteon menetelmiä samaan tulokseen päästään yhden otoksen tutkimisella. Ratkaisuun palataan tuonnempana.

Tilastollinen päättely

Tilastollisen päättelyn (inferential statistics) tavoitteena on **luotettavien johtopäätösten tekeminen perusjoukosta otoksen perusteella**. Tilastollisen päättelyn avulla pyritään arvioimaan, kuinka hyvin otoksesta saadut tulokset voidaan yleistää perusjoukon tuloksiksi.

Otoksesta määritetyt tunnusluvut ovat hyvin harvoin täysin samat kuin perusjoukon vastaavat tunnusluvut. Jos esimerkiksi mitataan sadan satunnaisesti valitun varusmiehen pituus ja lasketaan heidän pituuksiensa keskiarvo, se ei hyvin todennäköisesti ole millimetrin tarkkuudella sama kuin kaikkien sillä hetkellä palveluksessa olevien varusmiesten pituuksien keskiarvo. Toisaalta voidaan olla melko varmoja, että otoksen keskiarvo (sadan varusmiehen pituuksien keskiarvo) ei poikkea perusjoukon keskiarvosta (kaikkien varusmiesten pituuksien keskiarvo) montakaan senttimetriä.

Otoksesta määritetyn tunnusluvun ja perusjoukon vastaavan tunnusluvun välistä eroa sanotaan **otantavirheeksi**. Yleensä otantavirheen enimmäissuuruus voidaan arvioida.

Seuraavassa tutustutaan kahteen tilastollisen päättelyn keskeiseen asiaan: **parametrien estimointiin** ja **hypoteesien testaukseen**.

7.1 Parametrien estimointi

Estimointi

Estimoinnilla tarkoitetaan sitä tilastollisen päättelyn osa-aluetta, jossa otoksesta laskettujen tunnuslukujen avulla tehdään arvioita perusjoukon vastaavista parametreista. Useimmiten estimoidaan keskiarvoja tai prosenttiosuuksia.

ESIMERKKI 1

Tutkittiin, kuinka monta kertaa suomalaiset keskimäärin saunovat kuukaudessa. Asian selvittämiseksi poimittiin satunnaisesti 520 henkilöä, joilta kysyttiin saunomiskertojen lukumäärää. Keskiarvoksi saatiin 6,2 kertaa kuukaudessa. Tämä keskiarvo voidaan yleistää koskemaan kaikkia suomalaisia. Saatua lukua 6,2 sanotaan **piste-estimaatiksi**.

Tietenkin on selvää, ettei 520 henkilön otoksen perusteella voida täysin luotettavasti päätellä, että keskiarvo on juuri 6,2 eikä esimerkiksi 6,3. Onkin tärkeämpää pyrkiä löytämään rajat, joiden välillä todellinen kaikkien suomalaisten saunomiskertojen keskiarvo on. Nämä rajat riippuvat paitsi otoksen koosta myös muun muassa siitä, kuinka varmoja tuloksen oikeellisuudesta halutaan olla. Laskettua väliä sanotaan **väliestimaatiksi**.

Otoksen tunnusluvut ovat perusjoukon parametrien **estimaattoreita** (estimator) ja niiden arvot **estimaatteja** (estimate). Seuraavassa taulukossa luetellaan perusjoukon parametreja ja niitä vastaavia piste-estimaattoreita:

Perusjoukon parametri	Estimaattori
Odotusarvo μ	Keskiarvo $\bar{x}$
Keskihajonta σ	Otoskeskihajonta s
Varianssi σ^2	Varianssi s^2
Suhteellinen osuus p	Suhteellinen frekvenssi $\hat{p}$

Estimoitaessa perusjoukon parametreja lasketaan aluksi piste-estimaatti ja muodostetaan sen jälkeen väliestimaatti. Väliestimaattiin liittyviä rajoja sanotaan **luottamusrajoiksi** ja niiden määrittämää väliä **luottamusväliksi** (confidence interval).

Luottamusväli

Väliä [a, b] sanotaan parametrin t luottamusväliksi **luottamustasolla** (confidence level) $\mathbf{1-\alpha}$, jos

$P(a \leq t \leq b) = 1 - \alpha$.

Todennäköisyydellä $1-\alpha$ väli $[a, b]$ peittää perusjoukon parametrin t arvon.

Luottamusväli ilmaisee otoksesta laskettuun tunnuslukuun sisältyvän virhemarginaalin. Asian voi ilmaista myös siten, että luottamusväli määrittelee rajat, joiden väliin $100 \cdot (1-\alpha)$ % tunnusluvuista asettuu, jos tutkimusta jatketaan loputtomasti.

Aikaisemmin normaalijakauman käsittelyn yhteydessä todettiin, että

1. normaalijakaumaa $N(\mu,\sigma)$noudattavan satunnaismuuttujan X otoskeskiarvon jakauma noudattaa normaalijakaumaa $N(\mu, \frac{\sigma}{\sqrt{n}})$
2. kun n on riittävän suuri (vähintään 30), tavallisesti tuntematon perusjoukon keskihajonta σ voidaan korvata otoksesta lasketulla keskihajonnalla s.
3. 95 % normaalisti jakautuneen satunnaismuuttujan arvoista on välillä odotusarvo $\pm 1{,}96 \cdot$ keskihajonta.

Keskeisen raja-arvolauseen perusteella voidaan lisäksi päätellä, että satunnaismuuttujan X jakaumasta riippumatta otoskeskiarvon jakauma on likimäärin normaalijakauma, kun n on riittävän suuri ($n \geq 30$).

ESIMERKKI 2

Edellisen esimerkin otoksesta laskettu saunomiskertojen keskihajonta oli 2,4 kertaa kuukaudessa. Laske 95 %:n väliestimaatti saunomiskertojen lukumäärälle.

Otoksen kokoa 520 voidaan pitää riittävän suurena, joten keskeisen raja-arvolauseen mukaan saunomiskertojen keskiarvo

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}).$$

Perusjoukon keskihajonta voidaan korvata otoksen hajonnalla s (edellä kohta 2), joka oli siis 2,4.

Edellä kohdan 3 perusteella

$$P(6{,}2 - 1{,}96 \cdot \frac{2{,}4}{\sqrt{520}} \leq \mu \leq 6{,}2 + 1{,}96 \cdot \frac{2{,}4}{\sqrt{520}}) = 0{,}95$$

$$P(6{,}0 \leq \mu \leq 6{,}4) = 0{,}95.$$

Tämän tutkimuksen mukaan suomalaisten saunomiskertojen kuukausikeskiarvo vaihtelee lukujen 6,0 ja 6,4 välillä 95 %:n varmuudella.

Odotusarvon luottamusväli

Perusjoukon odotusarvon luottamusväli luottamustasolla $1-\alpha$ on

$$\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.$$

Kaavassa
$\bar{x}$ = otoskeskiarvo
n = otoksen koko ($n \geq 30$)
s = otoksen keskihajonta
$z_{\alpha/2}$ = luottamustasoon liittyvä kriittinen arvo.

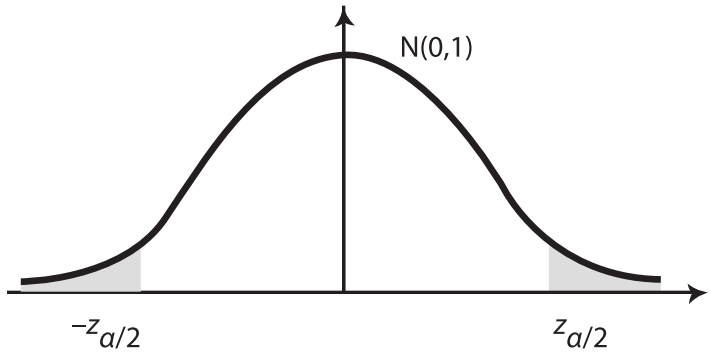
Luottamusväli on siis piste-estimaatti $\pm$ virhemarginaali.

Yleisimmin käytettävät luottamustasot ja niihin liittyvät kriittiset arvot ovat seuraavat

Luottamustaso	95 %	99 %	99,9 %
Kriittinen arvo $z_{\alpha/2}$	1,96	2,58	3,30

Luottamustaso osoittaa, kuinka todennäköistä on, että perusjoukkoa kuvaava tunnusluku on tietyllä luottamusvälillä.

Kriittisiä arvoja havainnollistaa seuraava kuvio:



ESIMERKKI 3

Vaatetehdas halusi selvittää otantatutkimuksen avulla, mikä on suomalaisten miesten keskipaino (kg). Otoksen kooksi suunniteltiin 800 miestä, joista tavoitettiin 752. Otokseen tulleiden keskipaino oli 78,3 kg ja painojen keskihajonta 22,4 kg. Määritä 99 %:n luottamusväli suomalaismiesten keskimääräiselle painolle.

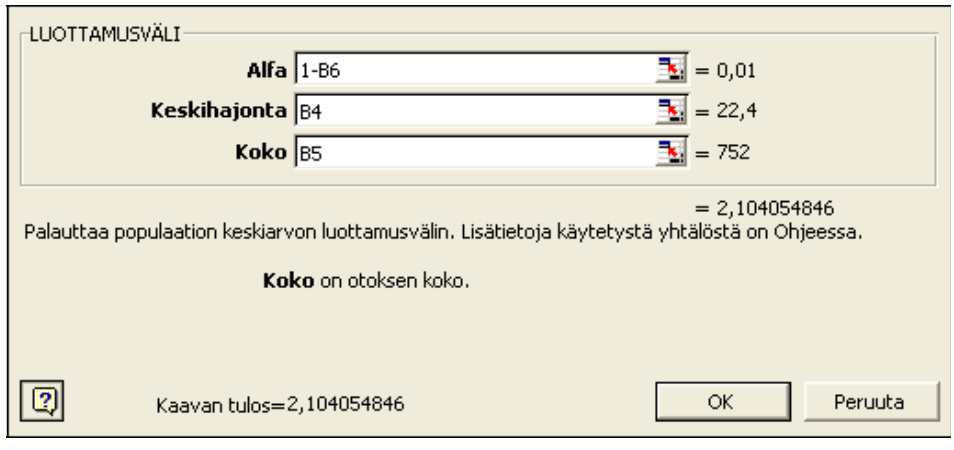
$\bar{x} = 78{,}3$ kg; $s = 22{,}4$ kg; $n = 752$ ja $z_{0{,}005} = 2{,}58$

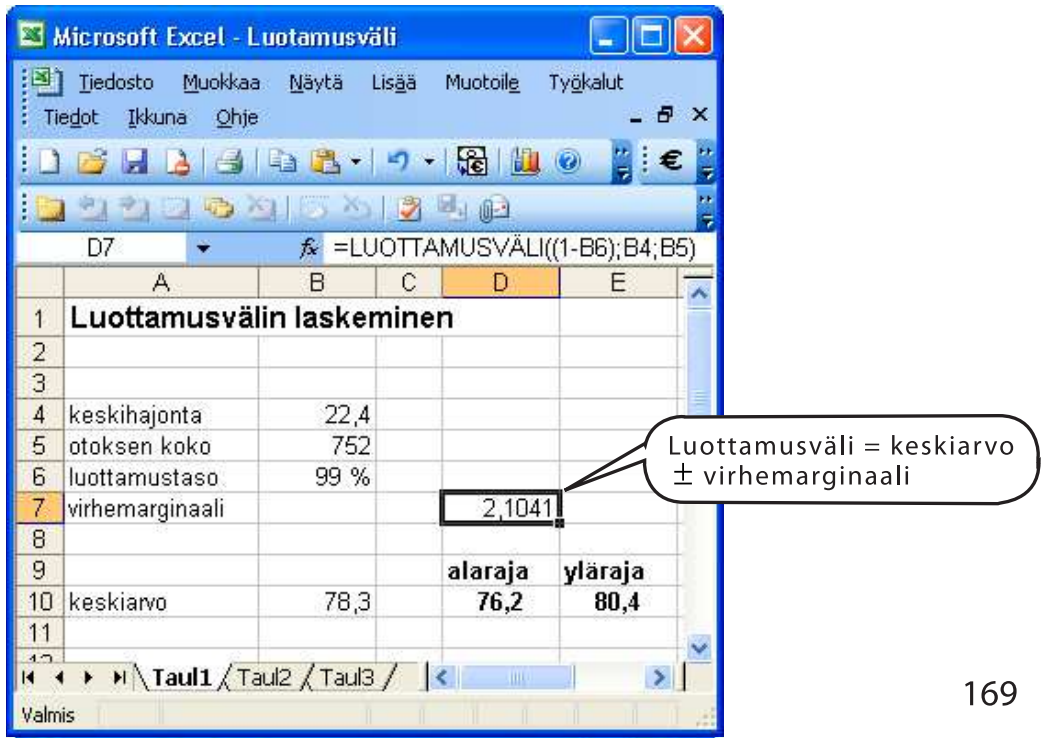
$$78{,}3 - 2{,}58 \cdot \frac{22{,}4}{\sqrt{752}} \leq \mu \leq 78{,}3 + 2{,}58 \cdot \frac{22{,}4}{\sqrt{752}}$$

$$76{,}2 \leq \mu \leq 80{,}4$$

Suomalaisten miesten keskimääräinen paino on tämän tutkimuksen mukaan 99 %:n varmuudella 76,2 kg:n ja 80,4 kg:n välillä.

Luottamusvälin pituus voidaan laskea myös Excelin funktion LUOTTAMUSVÄLI(α; s; n) (engl. CONFIDENCE) avulla seuraavasti:





Funktion LUOTTAMUSVÄLI avulla saadaan virhemarginaali, joka on 2,1041 kg (solu D7). Kun se vähennetään keskiarvosta, saadaan luottamusvälin alaraja. Vastaavasti yläraja saadaan, kun solun D7 arvo lisätään keskiarvoon.

Prosenttiosuuden luottamusväli

Tarkastellaan toistokoetta, jossa tapahtuma A voi esiintyä tai jäädä esiintymättä. Jos koe tehdään n kertaa, siihen liittyvä satunnaismuuttuja X on jakautunut Bin(n, p). Tassä X on tapahtuman A esiintymiskertojen lukumäärä ja p on tapahtuman A esiintymistodennäköisyys. Kun toistojen määrä on riittävän suuri, satunnaismuuttujan X jakauma on keskeisen raja-arvolauseen mukaan

$\mathrm{N}\left(np, \sqrt{np(1-p)}\right)$.

Muodostetaan uusi satunnaismuuttuja $P = \frac{X}{n}$, joka on tapahtuman A suhteellinen osuus n toistokokeessa. Voidaan osoittaa, että riittävän suurilla n:n arvoilla P:n jakauma on likimäärin $\mathrm{N}\left(p, \sqrt{\frac{p \cdot (1-p)}{n}}\right)$.

Kun jatketaan päättelyä samaan tapaan kuin odotusarvon luottamusväliä määritettäessä, saadaan seuraava tulos:

Suhteellisen osuuden p luottamusväli luottamustasolla $1-\alpha$ on

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}.$$

Kaavassa

$\hat{p}$ = otoksesta laskettu suhteellinen frekvenssi
n = otoksen koko
$z_{\alpha/2}$ = luottamustasoon liittyvä kriittinen arvo.

Kysymykseen, mistä n:n arvosta alkaen luottamusväli voidaan laskea edellä olevalla kaavalla, on vaikea antaa yksiselitteistä vastausta. Jos kuitenkin $p \approx 0{,}5$, voidaan kaavaa käyttää, kun $n \geq 30$. Vastaavasti jos $p \approx 0{,}2$, tulisi n:n olla vähintään 200. Jos $p \approx 0{,}05$, on n:n oltava noin 1 500.

ESIMERKKI 4

Tarkastellaan nyt tämän luvun alussa esitettyä ongelmaa, jossa vertailtiin kahden puolueen kannatusta. Siinä haastateltiin 1 016 äänioikeutettua suomalaista, joista 24,6 % ilmoitti kannattavansa puoluetta A ja 23,0 % puoluetta B. Tehtävänä oli selvittää, voidaanko tämän otantatutkimuksen tulos luotettavasti yleistää koskemaan kaikkia äänioikeutettuja eli onko puolueen A kannatus koko kansan keskuudessa suurempi kuin puolueen B kannatus.

Lasketaan molempien puolueiden kannatukselle 95 %:n luottamusvälit:

Puolue A

$\hat{p} = 0{,}246$

$1 - \hat{p} = 1 - 0{,}246 = 0{,}754$

$n = 1\,016$

$z_{0{,}025} = 1{,}96$

$$0{,}246 - 1.96 \cdot \sqrt{\frac{0{,}246 \cdot 0{,}754}{1\,016}} \leq p \leq 0{,}246 + 1{,}96 \cdot \sqrt{\frac{0{,}246 \cdot 0{,}754}{1\,016}}$$

$0{,}220 \leq p \leq 0{,}272$

Puolueen A kannatus vaihtelee siis 95 %:n varmuudella 22,0 %:n ja 27,2 %:n välillä.

Puolue B

$\hat{p} = 0{,}230$

$1 - \hat{p} = 1 - 0{,}230 = 0{,}770$

$n = 1\,016$

$z_{0{,}025} = 1{,}96$

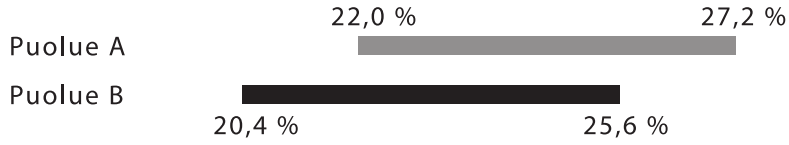
$$0{,}230 - 1{,}96 \cdot \sqrt{\frac{0{,}230 \cdot 0{,}770}{1\,016}} \leq p \leq 0{,}230 + 1{,}96 \cdot \sqrt{\frac{0{,}230 \cdot 0{,}770}{1\,016}}$$

$0{,}204 \leq p \leq 0{,}256$

Puolueen B kannatus puolestaan vaihtelee 95 %:n varmuudella 20,4 %:n ja 25,6 %:n välillä.

Ei siis voida olla varmoja, että puolueen A kannatus kaikkien äänioikeutettujen joukossa olisi suurempi kuin puolueen B kannatus.

Asiaa voidaan havainnollistaa esittämällä graafisesti molempien puolueiden kannatusta kuvaavat luottamusvälit:



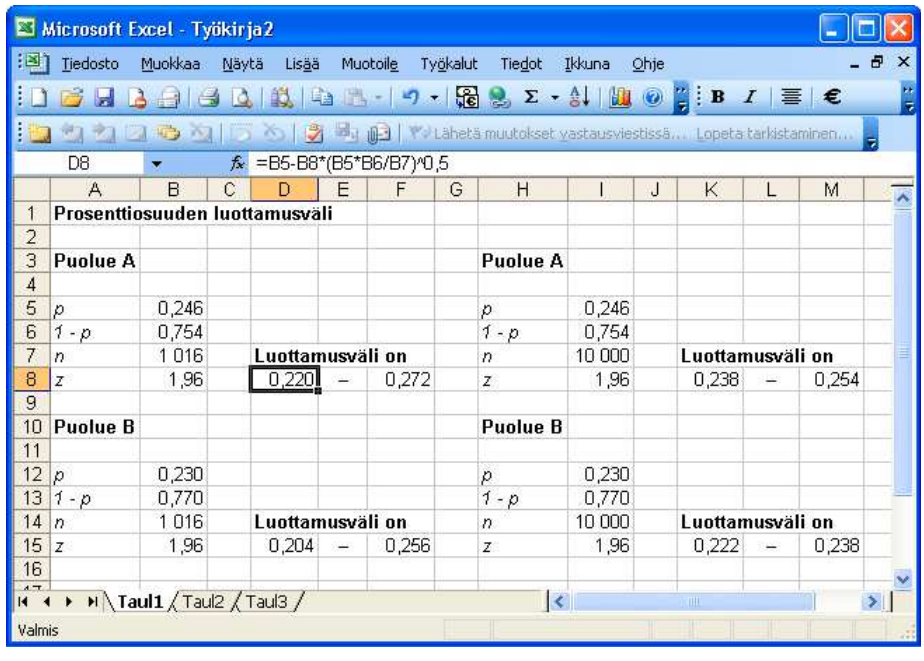
Koska luottamusvälit ovat osittain ”päällekkäin”, ei voida olla varmoja, että puolue A on puoluetta B suositumpi.

Mielenkiintoista on myös tutkia, kuinka iso otoksen pitäisi olla, jotta luotettavia johtopäätöksiä edellä mainittujen kannatuslukujen perusteella voitaisiin tehdä.

ESIMERKKI 5

Tutki taulukkolaskentaohjelman avulla, kuinka suuri otoksen koon edellisessä esimerkissä pitäisi olla, jotta puolueiden edellä olevien kannatuslukujen perusteella voitaisiin 95 %:n varmuudella päätellä, että puolue A on suositumpi kuin puolue B.

Syötetään tarvittava laskentakaava taulukkolaskentaohjelman soluihin ja kokeillaan eri *n*:n arvoilla.



Luottamusvälit on aluksi laskettu alkuperäisellä otoskoolla. Muuttamalla *n*:n arvoa havaitaan, että otoksen koon olisi pitänyt olla noin 10 000 henkilöä, jotta tutkimuksessa saadun puolueiden kannatuseron perusteella voitaisiin varmistua, että puolue A on puoluetta B suositumpi kaikkien äänestäjien joukossa.

Edellä kokeilemalla saatu otoskoko voidaan laskea yksinkertaisen kaavan avulla.

Oletetaan, että piste-estimaatin $\hat{p}$ enimmäisvirheeksi luottamustasolla $1-\alpha$ sallitaan $\pm e$ prosenttiyksikköä. Toisin sanottuna perusjoukon parametrin on oltava todennäköisyydellä $1-\alpha$ välillä $\hat{p} \pm e$. Mikä otoskoon on vähintään oltava, jotta kyseiseen tulokseen päästäisiin?

Otoskoko

Otoksen vähimmäiskoko voidaan laskea kaavalla

$$n = z_{\alpha/2}{}^2 \cdot \frac{\hat{p} \cdot (1 - \hat{p})}{e^2}, \text{ jossa}$$

n = otoksen koko
$\hat{p}$ = otoksesta laskettu piste-estimaatti
e = sallittu enimmäisvirhe desimaalilukuna
$z_{\alpha/2}$ = luottamustasoon liittyvä kriittinen arvo.

ESIMERKKI 6

Tuotteen markkinaosuus on noin 20 %. Asiaa halutaan tutkia tarkemmin markkinatutkimuksessa. Kuinka suuri otoksen koon on vähintään oltava, jotta saadun tuloksen enimmäisvirhe 95 %:n varmuudella olisi enintään ±2 prosenttiyksikköä?

$\hat{p} = 0{,}2;$ $1 - \hat{p} = 0{,}8;$ $e = 0{,}02;$ $z_{0{,}025} = 1{,}96$

$$n = 1{,}96^2 \cdot \frac{0{,}2 \cdot 0{,}8}{0{,}02^2} = 1\,537$$

Otoksen koon olisi oltava vähintään 1 500 tilastoyksikköä, jotta haluttuun tarkkuuteen päästäisiin.

Usein halutaan selvittää, kuinka suuri on enimmäisvirhe prosenttiyksiköinä, kun otoksen koko tunnetaan.

Kun edellisestä kaavasta ratkaistaan e^2, saadaan $e^2 = z_{\alpha/2}{}^2 \cdot \frac{\hat{p} \cdot (1 - \hat{p})}{n}$, josta saadaan lopullinen kaava.

Virhemarginaali

Kun estimoidaan perusjoukon prosenttiosuutta, piste-estimaatin enimmäisvirhe desimaalilukuna saadaan kaavasta

$$e = \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

ESIMERKKI 7

Seurakunta halusi tutkia, onko naispappeuden vastustus vähentynyt täysi-ikäisten seurakuntalaisten joukossa. Tutkimusta varten poimittiin 18 vuotta täyttäneistä seurakunnan jäsenistä 946 henkilön otos. Heistä 8,6 % vastusti naispappeutta. Voidaanko 95 prosentin varmuudella sanoa, että naispappeutta vastustaa vähemmän kuin 10 % seurakunnan täysi-ikäisistä jäsenistä?

$n = 946;\ \hat{p} = 0{,}086;\ 1 - \hat{p} = 1 - 0{,}086 = 0{,}914;\ z_{0{,}025} = 1{,}96$

$$e = \pm 1{,}96 \cdot \sqrt{\frac{0{,}086 \cdot 0{,}914}{946}} \approx \pm 0{,}018$$

Piste-estimaatin enimmäisvirhe on siis 1,8 prosenttiyksikköä molempiin suuntiin. Koska 8,6 % + 1,8 %-yks. = 10,4 %, ei voida olla 95-prosenttisen varmoja siitä, että naispappeuden vastustajia seurakuntalaisten joukossa on vähemmän kuin 10 %.

7.2 Hypoteesien testaus

Lähes kaikkeen tutkimustoimintaan kuuluu tiettyjen olettamusten tai väitteiden todenperäisyyden testaaminen. Tilastollisten testien avulla tutkitaan perusjoukosta tehtyjen olettamusten eli **hypoteesien** paikkansapitävyyttä.

ESIMERKKI 8

Sähkölamppuja valmistavan teollisuuslaitoksen laaduntarkkailussa on pitkäaikaisten tilastojen perusteella todettu, että yrityksen valmistamien sähkölamppujen keskimääräinen palamisaika on 1 200 tuntia ja palamisaikojen keskihajonta 100 tuntia.

Yritys tarkkailee tuotteen laatua säännöllisesti. Viimeksi tehdyssä laaduntarkastuksessa todettiin, että 100:n satunnaisesti valitun lampun keskimääräinen palamisaika oli 1 147 tuntia ja palamisaikojen keskihajonta 96 tuntia. Voidaanko päätellä, että lamppujen laatu on huonontunut?

On selvitettävä, johtuuko otokseen valittujen lamppujen keskimääräisen palamisajan lyheneminen aikaisempaan verrattuna **sattumasta** vai onko laatu todella huonontunut. Jos päädytään jälkimmäiseen johtopäätökseen, on tietenkin ryhdyttävä selvittämään syitä tähän, jotta tuotannon laatu saataisiin vähintään entiselle tasolle.

Suorittamalla **tilastollinen testi** otoksesta laskettujen tunnuslukujen avulla voidaan varmistaa, onko lamppujen laatu huonontunut.

ESIMERKKI 9

Lääkärit ovat työssään havainneet, että aspiriinilla saattaa olla sydäninfarkteja ehkäisevä vaikutus. Asian selvittämiseksi tehdään tutkimus, jossa tutkitaan kahta vaihtoehtoa. Toista näistä vaihtoehdoista nimitetään **nollahypoteesiksi H_0** (null hypothesis) ja toista **vaihtoehtoiseksi hypoteesiksi H_1** (alternative hypothesis). Yleensä nollahypoteesi pyritään hylkäämään, jolloin vaihtoehtoinen hypoteesi tulee voimaan. Tässä tapauksessa hypoteesit olisivat seuraavat:

H_0: Aspiriinilla ei ole sydäninfarkteja ehkäisevää vaikutusta.

H_1: Aspiriini ehkäisee sydäninfarktien syntymistä.

Siihen, millä tavalla edellinen tutkimus käytännössä toteutetaan, ei tässä yhteydessä puututa.

Käytännössä nollahypoteesi H_0 on perusolettamus tai vallitseva käsitys siitä, kuinka asia on. Nollahypoteesia pidetään totena, kunnes toisin pystytään osoittamaan. Jos löydetään todisteet nollahypoteesia vastaan, se hylätään ja vaihtoehtoinen hypoteesi H_1 tulee voimaan.

Tilastollinen testaus voidaan jakaa neljään päävaiheeseen:

1. Asetetaan nollahypoteesi ja vaihtoehtoinen hypoteesi.
2. Poimitaan perusjoukosta yksi tai useampia otoksia.
3. Tehdään tilanteeseen sopiva tilastollinen testi.
4. Tulkitaan testin tulos.

Tilastollisen päättelyn virheet

Hypoteesien testauksen jälkeenkään ei voida olla täysin varmoja, että perusjoukkoa koskeva johtopäätös on oikea. On neljä mahdollisuutta tehdä päätös:

1. H_0 hylätään, kun se on epätosi.
2. H_0 hylätään, vaikka se on tosi.
3. H_0 jää voimaan, vaikka se on epätosi.
4. H_0 jää voimaan, kun se on tosi.

Näistä ensimmäinen ja neljäs ovat oikeita ja toinen ja kolmas vääriä johtopäätöksiä.

Jos nollahypoteesi H_0 hylätään, vaikka se on tosi, tehdään **1. lajin virhe** eli **hylkäämisvirhe**.

Jos taas H_0 jää voimaan, vaikka se pitäisi hylätä, tehdään **2. lajin virhe** eli **hyväksymisvirhe**.

Sama voidaan esittää taulukkona:

	Todellisuus	
Päätös	**H_0 epätosi**	**H_0 tosi**
H_0 epätosi	1 Oikea päätös	2 Hylkäämisvirhe
H_0 tosi	3 Hyväksymisvirhe	4 Oikea päätös

Merkitsevyystaso = hylkäämisvirheen todennäköisyys

Nollahypoteesia testattaessa päätetään etukäteen, kuinka suuri hylkäämisvirheen riski otetaan. Hylkäämisvirheen tekemisen todennäköisyyttä α sanotaan **merkitsevyystasoksi**. Tilasto-ohjelmat laskevat automaattisesti hylkäämisvirheen todennäköisyyden, jota sanotaan ***p*-arvoksi** (myös Significance tai Sig.). Saatu *p*-arvo osoittaa, kuinka suuri on väärän johtopäätöksen todennäköisyys, jos nollahypoteesi hylätään, eli kuinka suurella todennäköisyydellä vaihtoehtoinen hypoteesi on väärä.

Käytetyimmät merkitsevyystasot (hylkäämisvirheen todennäköisyydet) ovat seuraavat:

1. $\alpha \leq 0,001$ — Tulos tai ero on **tilastollisesti erittäin merkitsevä**.
2. $0,001 < \alpha \leq 0,01$ — Tulos tai ero on **tilastollisesti merkitsevä**.
3. $0,01 < \alpha \leq 0,05$ — Tulos tai ero on **tilastollisesti melkein merkitsevä**.

Huomautus

Tuloksen tilastollinen merkitsevyys ei välttämättä tarkoita sitä, että sillä olisi käytännön merkitystä.

ESIMERKKI 10

Oletetaan, että edellisen esimerkin tutkimuksessa valittiin merkitsevyystasoksi 1 %. Jos nollahypoteesi tulisi tutkimuksen perusteella hylätyksi, voitaisiin päätellä, että aspiriinilla todennäköisesti on sydäninfarkteja ehkäisevä vaikutus. Väärän johtopäätöksen todennäköisyys on enintään 1 %.

Tilastollinen testi voi olla **yksi**- tai **kaksisuuntainen**.

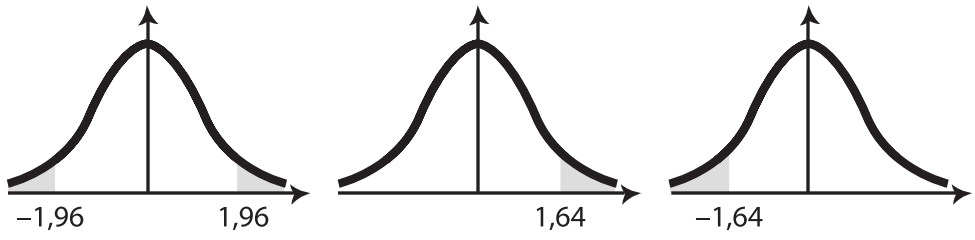
ESIMERKKI 11

Kun testataan, onko sähkölamppujen keskimääräinen palamisaika 1 200 tuntia, voidaan nollahypoteesiksi valita

$H_0: \mu = 1\,200$. Kaksisuuntaisessa testissä vaihtoehtoinen hypoteesi on $\mathbf{H_1: \mu \neq 1\,200}$. Yksisuuntaisessa testissä vaihtoehtoinen hypoteesi voidaan asettaa kahdella tavalla: $\mathbf{H_1: \mu > 1\,200}$ tai $\mathbf{H_1: \mu < 1\,200}$.

Nollahypoteesin hylkääminen perustuu testimuuttujan saamaan arvoon. Se hylätään, jos **testimuuttujan arvo** sijaitsee **hylkäämisalueella**.

Hylkäämisalue ja vastaavat hypoteesit normaalijakaumaan perustuvassa testauksessa 5 %:n merkitsevyystasolla ovat seuraavat (hylkäämisalueet varjostettu):



$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$

$H_0 : \mu = \mu_0$
$H_1 : \mu > \mu_0$

$H_0 : \mu = \mu_0$
$H_1 : \mu < \mu_0$

7.3 Parametriset testit

Tilastollisia testejä on kehitetty erittäin runsaasti. Testiä valittaessa on varmistuttava, että testin käytön edellytykset ovat voimassa. Aluksi päätetään, käytetäänkö **parametrista** vai **jakaumasta riippumatonta** testiä.

Käsitellään aluksi parametrisia testejä. Niiden käyttöön liittyy oletuksia perusjoukon tunnusluvuista ja jakaumasta. Testejä käytettäessä on tarkasteltavan muuttujan jakauma perusjoukossa käytännössä tunnettava ainakin likimain. Lisäksi testattavan muuttujan mitta-asteikkona on oltava **välimatka-** tai **suhdeasteikko**.

Yleisimpiä parametrisia testejä ovat **keskiarvotestit**. Tarkastellaan normaalijakaumaan perustuvaa yksisuuntaista testiä.

ESIMERKKI 12

Palataan esimerkkiin 8, jossa sähkölamppuja valmistavan teollisuuslaitoksen laaduntarkkailu oli pitkäaikaisten tilastojen perusteella todennut, että sähkölamppujen keskimääräinen palamisaika on 1 200 tuntia ja palamisaikojen keskihajonta 100 tuntia. Viimeksi tehdyssä tarkastuksessa 100:n satunnaisesti valitun lampun keskimääräinen palamisaika oli kuitenkin 1 147 tuntia ja keskihajonta 96 tuntia. Tuotantopäällikkö arvelee, että prosessissa on ollut häiriö. Asian todentamiseksi tehdään tilastollinen testi.

Asetetaan nollahypoteesi $\mathbf{H_0: \mu = 1\,200}$.

Jos nollahypoteesi jää voimaan, tämän otoksen perusteella ei ilmeisesti voida päätellä, että lamppujen laatu olisi huonontunut. Otoksessa todettu vaihtelu johtuu vain sattumasta.

Vastaavasti vaihtoehtoinen hypoteesi on $\mathbf{H_1: \mu < 1\,200}$.

Jos H_0 hylätään, päätellään, että lamppujen laatu on heikentynyt.

Valitaan merkitsevyystasoksi 0,1 %.

Kuten luvussa 6 sivulla 144 on esitetty, otoskeskiarvon $\overline{x}$ jakauma on likimäärin $N\left(\mu, \frac{s}{\sqrt{n}}\right)$, kun $n \geq 30$.

Tällöin normitettu satunnaismuuttuja $Z = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$ noudattaa jakaumaa N (0, 1).

Muuttujaa Z käytetään testimuuttujana. Sen kriittiset arvot saadaan yhtälöstä $P(Z < z_\alpha) = \alpha$.

Koska $\alpha = 0{,}1\ \% = 0{,}001$, saadaan

$P(Z < -z_{0{,}001}) = 0{,}001$ eli $-z_{0{,}001} = -3{,}09$ (ks. liitteenä olevaa taulukkoa Normitetun normaalijakauman kertymäfunktion arvoja).

Jos testimuuttujan arvo on pienempi kuin –3,09, nollahypoteesi hylätään.

Koska nollahypoteesin mukaan $\mu = 1\,200$ ja 100:n kappaleen otoksesta lasketut tunnusluvut ovat $\overline{x} = 1\,147$ ja $s = 96$, testimuuttujan arvoksi saadaan

$$Z = \frac{1147 - 1200}{\frac{96}{\sqrt{100}}} = -5{,}52 < -3{,}09.$$

Tulkinta: Testimuuttujan arvo on hylkäämisalueella, joten nollahypoteesi hylätään. Lamppujen laatu on syystä tai toisesta huonontunut. Tulos on tilastollisesti erittäin merkitsevä.

ESIMERKKI 13

Erään hajuvesimerkin markkinaosuus oli aikaisempina vuosina ollut maassamme keskimäärin 14,0 %. Tutkimuksella haluttiin selvittää, oliko markkinaosuus viimeisen vuoden aikana muuttunut. Tutkimusta varten poimittiin eri puolilta maata yhteensä 40:n hajuvesiä myyvän liikkeen otos. Kuukauden aikana tapahtuneen myynninseurannan perusteella hajuvesimerkin markkinaosuus (%) otokseen poimituissa liikkeissä oli seuraava:

12,5	15,7	14,1	11,9	14,8	12,9	13,6	15,4	12,0
14,2	13,9	11,3	15,0	11,8	13,3	15,2	12,7	14,3
15,2	11,9	13,7	14,2	11,7	14,4	13,0	12,9	14,8
15,0	13,3	14,0	14,4	12,9	14,1	12,5	13,3	12,8
14,5	13,2	15,1	11,9					

Voidaanko päätellä, että hajuvesimerkin keskimääräinen markkinaosuus on muuttunut aikaisemmista vuosista?

Tutkitaan asiaa sekä 5 %:n että 1 %:n merkitsevyystasoilla. Asetetaan hypoteesit $H_0 : \mu = 14\ \%; H_1 : \mu \neq 14\ \%$ ja ratkaistaan ongelma tilasto-ohjelman avulla:

Hypothesis Test: Mean vs. Hypothesized Value	
14,0000	hypothesized value
13,5850	mean Group 1
1,1915	std. dev.
0,1884	std. error
40	n
–2,20	z
,0138	p-value (one-tailed)
,0276	p-value (two-tailed)

Tämä ohjelma tulostaa nollahypoteesiksi asetetun arvon (14,0000), otoskeskiarvon (13,5850), otoksen keskihajonnan (1,1915), otoskeskiarvon keskivirheen (0,1884), otoksen koon (40), testimuuttujan Z arvon (–2,20) sekä p-arvon yksisuuntaisessa testissä (0,0138) ja kaksisuuntaisessa testissä (0,0276).

Johtopäätökset: Koska testimuuttujan arvo $-2{,}20 < -1{,}96 = -z_{0{,}025}$, nollahypoteesi hylätään 5 %:n merkitsevyystasolla.

Toisaalta, koska $-2{,}20 > -2{,}58 = z_{0{,}005}$, nollahypoteesia ei hylätä 1 %:n merkitsevyystasolla vaan se jää voimaan.

Tulkinta: Hajuvesimerkin markkinaosuus on 95 %:n varmuudella muuttunut 14 %:sta, mutta 99 %:n varmuutta asiasta ei ole. Tulos, jonka mukaan markkinaosuus on muuttunut, on siten tilastollisesti melkein merkitsevä. Väärän johtopäätöksen riski on puolestaan 2,76 % (p-arvo).

Keskiarvotesti

Perusjoukkoon liittyvän muuttujan X odotusarvon testaamiseksi poimitaan n ($n \geq 30$) tilastoyksikön otos kyseisestä perusjoukosta. Otoksen perusteella laskettu keskiarvo olkoon $\overline{x}$ ja keskihajonta s. Asetetaan hypoteesit H_0: Perusjoukon odotusarvo $\mu = \mu_0$ ja H_1: $\mu \neq \mu_0$. Tällöin testimuuttuja $Z = \dfrac{\overline{x} - \mu_0}{\dfrac{s}{\sqrt{n}}} \sim N(0,1)$.

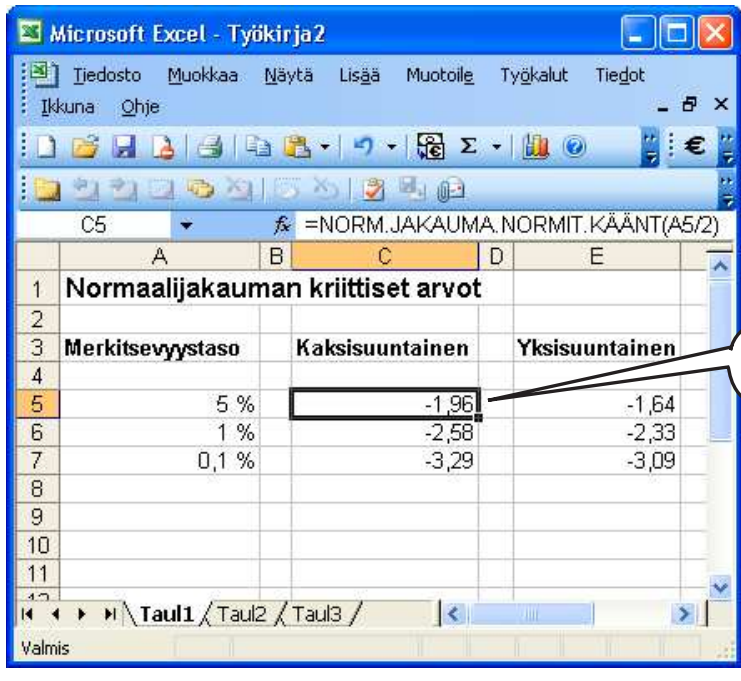
Nollahypoteesi H_0: $\mu = \mu_0$ hylätään ja vaihtoehtoinen hypoteesi H_1: $\mu \neq \mu_0$ tulee voimaan

1. 5 %:n merkitsevyystasolla, jos $Z < -1{,}96$ tai $Z > 1{,}96$
2. 1 %:n merkitsevyystasolla, jos $Z < -2{,}58$ tai $Z > 2{,}58$
3. 0,1 %:n merkitsevyystasolla, jos $Z < -3{,}30$ tai $Z > 3{,}30$.

Yksisuuntaisessa testissä nollahypoteesi H_0: $\mu = \mu_0$ hylätään ja H_1: $\mu < \mu_0$ ($\mu > \mu_0$) tulee voimaan

1. 5 %:n merkitsevyystasolla, jos $Z < -1{,}64$ ($Z > 1{,}64$)
2. 1 %:n merkitsevyystasolla, jos $Z < -2{,}33$ ($Z > 2{,}33$)
3. 0,1 %:n merkitsevyystasolla, jos $Z < -3{,}09$ ($Z > 3{,}09$).

Edellä olevat normaalijakauman kriittiset arvot saadaan myös Excelin funktion NORM.JAKAUMA.NORMIT.KÄÄNT(α) (engl. NORMSINV) avulla. Kaksisuuntaisessa testissä esimerkiksi 5 %:n merkitsevyystasolla argumentin α arvoksi annetaan 0,025. Seuraavassa kaikki edellä olevat arvot on laskettu Excelillä:



Kaksisuuntaisessa testissä 0,1 %:n merkitsevyystasolla arvoissa on yhden sadasosan ero, joka johtuu erilaisista laskentatarkkuuksista.

Otoskeskiarvon $\bar{x}$ jakauma voi poiketa normaalijakaumasta, jos $n < 30$, sitä enemmän, mitä pienempi n on. Pienten otosten keskiarvotestinä ei siten voida käyttää normaalijakaumaan perustuvaa testiä. Jos otos on kuitenkin poimittu likimain normaalisti jakautuneesta perusjoukosta, voidaan osoittaa, että testimuuttuja noudattaa Studentin t-jakaumaa vapausastein $n-1$. Studentin t-jakaumaan liittyvät kriittiset arvot on myös taulukoitu (ks. liitettä Studentin t-jakaumaan liittyvät kriittiset arvot). Studentin jakauman kriittiset arvot voidaan laskea helposti myös Excelin funktion TJAKAUMA.KÄÄNT (todennäköisyys; vapausasteet)(engl. TINV) avulla.

Studentin t-testi

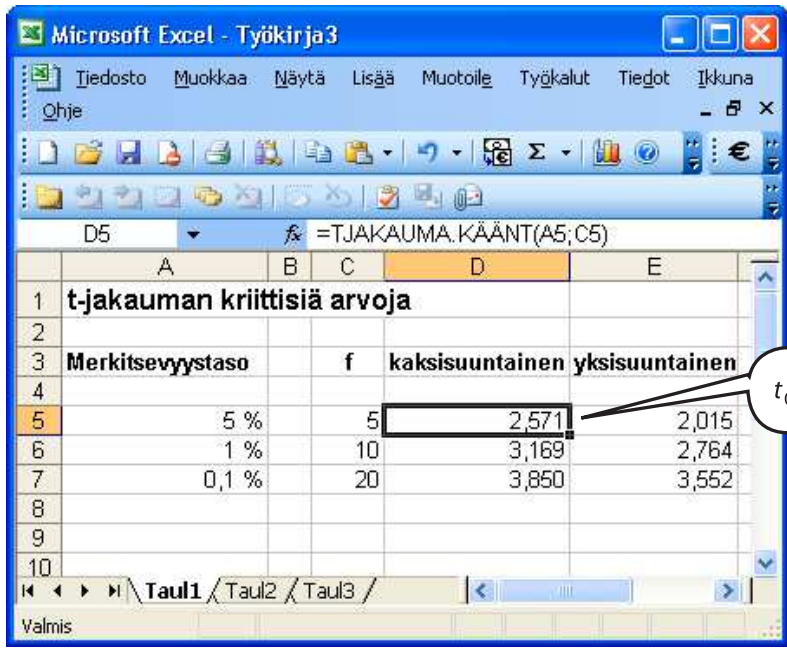
Poimitaan likimain normaalisti jakautuneesta perusjoukosta n tilastoyksikön satunnaisotos. Asetetaan hypoteesit H_0: $\mu = \mu_0$, H_1: $\mu \neq \mu_1$.

Testimuuttuja $t = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}}$ noudattaa t-jakaumaa vapausastein $f = n - 1$, jos nollahypoteesi on tosi.

Tilasto-ohjelmissa vapausasteita merkitään usein df:llä (degrees of freedom).

Nollahypoteesi H_0 hylätään ja vaihtoehtoinen hypoteesi H_1 tulee voimaan merkitsevyystasolla α, jos testimuuttujan arvo $t < -t_{\alpha/2}$ tai $t > t_{\alpha/2}$.

Seuraavaan Excel-taulukkoon on laskettu eräitä *t*-jakauman kriittisiä arvoja eri vapausastein:



Funktio laskee kaksisuuntaiseen *t*-jakaumaan liittyvän todennäköisyyden arvon. Solussa D5 on siis arvo $t_{0,025}$ vapausastein $f = 5$.

ESIMERKKI 14

Valmistaja ilmoittaa pesujauhepakkauksen nettopainoksi keskimäärin 960 g. Kuluttajatutkimuksessa punnittiin 8 eri liikkeestä ostetun pesujauhepaketin nettopaino. Keskimääräiseksi nettopainoksi saatiin 945 g ja painojen keskihajonnaksi 49 g. Testaa 5 %:n merkitsevyystasolla, onko valmistajan ilmoitus oikea.

Voidaan olettaa, että pesujauhepakkausten paino noudattaa likimäärin normaalijakaumaa, koska pakkaukset täytetään koneellisesti. Tällöin mittaustarkkuuden jakauma on normaali, ellei koneessa ole häiriötä.

Asetetaan hypoteesit H_0: $\mu = 960$ g; H_1: $\mu \neq 960$ g.

Testimuuttuja $t = \dfrac{945 - 960}{\dfrac{49}{\sqrt{8}}} = -0,866$

noudattaa Studentin *t*-jakaumaa vapausastein $f = 7$, jos nollahypoteesi on tosi.

Johtopäätökset: Kaksisuuntaisessa testissä arvolla $f = 7$ merkitsevyystasolla 5 % *t*-jakauman kriittinen arvo on $t_{0,025} = 2,365$. Jakauman symmetrisyyden perusteella $-t_{0,025} = -2,365$. Koska $-0,866 > -2,365$, nollahypoteesi jää voimaan 5 %:n merkitsevyystasolla.

Tulkinta: Valmistajan ilmoitusta ei testin perusteella voida kumota.

ESIMERKKI 15

Kaupan keskusliike testaa esimiesharjoittelijoiksi pyrkivät. Pitkän ajan tilastojen mukaan pyrkijöiden testissä saamien pistemäärien keskiarvo on ollut 52,0 pistettä. Eräänä vuonna testattiin 15 harjoittelijaksi pyrkivää. He saivat seuraavat pistemäärät: 64, 58, 69, 48, 55, 62, 57, 51, 46, 60, 52, 67, 53, 66 ja 58.

Testaa 1 %:n merkitsevyystasolla hypoteesi H_0: Pyrkijöiden taso ei ole noussut.

Koska tutkimuskohteena on ihmisten ominaisuudet, voidaan olettaa, että testipistemäärä perusjoukossa noudattaa likimäärin normaalijakaumaa.

Asetetaan hypoteesit H_0: $\mu = 52{,}0$; H_1: $\mu > 52{,}0$.

Tehdään testi tilasto-ohjelman avulla:

Hypothesis Test: Mean vs. Hypothesized Value	
52,000	hypothesized value
57,733	mean Group 1
6,995	std. dev.
1,806	std. error
15	n
14	df
3,17	t
,0034	p-value (one-tailed)

Johtopäätökset: Yksisuuntaisessa testissä vapausastein $f = 14$ kriittinen arvo $t_{0,01} = 2{,}624$. Koska $3{,}17 > 2{,}624$, nollahypoteesi H_0 hylätään. Väärän johtopäätöksen todennäköisyys on 0,3 % (p-arvo on 0,0034).

Tulkinta: Pyrkijöiden taso on tutkimuksen perusteella edellisiä vuosia parempi. Tulos on tilastollisesti merkitsevä.

Usein halutaan tutkia kahden toisistaan riippumattoman ryhmän välisiä eroja. Tällöin joudutaan testaamaan, onko ryhmien välillä perusjoukossa eroja. Aluksi tarkastellaan riippumattomia ryhmiä.

Kahden riippumattoman otoksen keskiarvotesti

Tutkitaan yhtä aikaa kahta riippumatonta otosta, joiden keskiarvot ja hajonnat tunnetaan. Jos otokset ovat riittävän suuria ($n \geq 30$), keskiarvojen erotusta voidaan testata normaalijakaumaan perustuvan testin avulla.

Asetetaan nollahypoteesi, että perusjoukoissa keskiarvot ovat yhtä suuret eli H_0: $\mu_1 = \mu_2$.

Testisuure $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)}}$ noudattaa tällöin likimain normitettua

normaalijakaumaa N(0, 1), jos nollahypoteesi on tosi.

Huomaa erityisesti, että perusjoukoista poimittujen otosten kokojen n_1 ja n_2 ei tarvitse olla yhtä suuret.

ESIMERKKI 16

Poimitaan 50 ja 65 kappaleen riippumattomat otokset kahdesta eri perusjoukosta. Otoksista lasketut keskiarvot ovat $\bar{x}_1 = 42{,}3$ ja $\bar{x}_2 = 47{,}5$. Vastaavasti otosten keskihajonnat ovat $s_1 = 10{,}2$ ja $s_2 = 8{,}8$. Testaa 1 %:n merkitsevyystasolla hypoteesi H_0: $\mu_1 = \mu_2$. Vaihtoehtoinen hypoteesi on $\mu_1 < \mu_2$.

Normaalijakauman kriittinen arvo yksisuuntaisessa testissä 1 %:n merkitsevyystasolla on –2,33. Lasketaan testisuureen arvo ja verrataan sitä kriittiseen arvoon.

$$Z = \frac{42{,}3 - 47{,}5}{\sqrt{\left(\dfrac{10{,}2^2}{50} + \dfrac{8{,}8^2}{65}\right)}} = -2{,}87.$$

Johtopäätökset: Koska –2,87 < –2,33, nollahypoteesi hylätään. Perusjoukon 1 keskiarvo on pienempi kuin perusjoukon 2 keskiarvo. Tulos on tilastollisesti merkitsevä.

ESIMERKKI 17

Työsuojelupiiri tutki työntekijöiden sairauspoissaoloja. Tarkoituksena oli selvittää, poikkeavatko teollisuuden ja palvelualojen työntekijöiden sairauspoissaolot toisistaan. Tutkimusta varten poimittiin paikkakunnan teollisuusyrityksistä 35 työntekijän ja palvelualojen yrityksistä 32 työntekijän satunnaisotokset. Otokseen poimittujen työntekijöiden sairauspoissaolot (päivää vuodessa) olivat seuraavat:

Teollisuus: 3, 15, 4, 22, 6, 10, 2, 0, 24, 18, 9, 8, 10, 10, 12, 0, 5, 21, 19, 14, 3, 1,13, 6, 11, 4, 12, 15, 10, 5, 13, 27, 12, 8, 9

Palvelualat: 9, 4, 16, 2, 12, 3, 8, 9, 15, 0, 4, 8, 8, 10, 12, 3, 6, 17, 14, 5, 9, 0, 3, 11,17, 18, 14, 15, 9, 4, 9, 17.

Voidaanko tutkimuksen perusteella päätellä, että teollisuustyöntekijöiden ja palvelualan työntekijöiden keskimääräiset sairauspoissaolot poikkeavat toisistaan?

Käytetään kaksisuuntaista testiä.

Asetetaan nollahypoteesi H_0: $\mu_1 = \mu_2$. Tällöin vaihtoehtoinen hypoteesi on H_1: $\mu_1 \neq \mu_2$.

Ratkaistaan tehtävä tietokoneohjelman avulla:

Hypothesis Test: Independent Groups		
Teollisuus	Palvelualat	
10,314	9,094	mean
6,850	5,354	std. dev.
35	32	n

1,221	difference (Teollisuus - Palvelualat)
1,496	standard error of difference
0,000	hypothesized difference
0,82	z
,4144	p-value (two-tailed)

Johtopäätökset: Koska testimuuttujan arvo 0,82 < 1,96, nollahypoteesia ei voida hylätä edes 5 %:n merkitsevyystasolla. Ohjelman tulostama *p*-arvo 0,4144 osoittaa, että erehtymisriski olisi yli 40 %, jos nollahypoteesi hylättäisiin.

Tulkinta: Tutkimuksen perusteella ei voida päätellä, että teollisuustyöntekijöiden ja palvelualojen työntekijöiden sairauspoissaolot poikkeaisivat toisistaan.

Normaalijakaumaan perustuvaa testausta ei edellisen esimerkin kaltaisissa tapauksissa voida luotettavasti käyttää, jos otosten koot ovat pieniä. Jos kuitenkin satunnaismuuttujien jakauma perusjoukoissa on likimain normaalinen, voidaan käyttää *t*-testiä.

Kahden riippumattoman otoksen keskiarvojen vertailu t-testin avulla

Tutkitaan satunnaismuuttujaa X kahdessa eri perusjoukossa. Tehdään seuraavat oletukset:

1. Satunnaismuuttujan X jakaumat perusjoukoissa ovat $N(\mu_1, \sigma_1)$ ja $N(\mu_2, \sigma_2)$.
2. Satunnaismuuttujan X keskihajonnat molemmissa perusjoukoissa ovat samat eli $\sigma_1 = \sigma_2$.
3. Molemmista perusjoukoista poimitaan otokset, joiden koko on n_1 ja n_2 tilastoyksikköä.

Asetetaan hypoteesi H_0: Perusjoukon keskiarvo molemmissa otoksissa on sama eli $\mu_1 = \mu_2$.

Testisuure $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s \cdot \sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$,

jossa $s^2 = \dfrac{(n_1 - 1) \cdot s_1^{\,2} + (n_2 - 1) \cdot s_2^{\,2}}{(n_1 - 1) + (n_2 - 1)}$

noudattaa *t*-jakaumaa vapausastein $f = n_1 + n_2 - 2$, jos nollahypoteesi on tosi.

ESIMERKKI 18

Tutkimuksessa haluttiin selvittää, poikkeavatko itä- ja länsisuomalaisten miesten pituudet toisistaan. Itäsuomalaisia poimittiin otokseen 22 ja länsisuomalaisia 28. Itäsuomalaisten miesten pituuksien keskiarvo oli 174,3 cm ja pituuksien keskihajonta 13,7 cm. Vastaavat tunnusluvut toisesta otoksesta olivat 176,2 cm ja 13,9 cm. Testaa 1 %:n merkitsevyystasolla hypoteesi, jonka mukaan itä- ja länsisuomalaiset ovat keskimäärin yhtä pitkiä. Vaihtoehtoinen hypoteesi on, että länsisuomalaiset miehet ovat keskimäärin pidempiä.

Voidaan katsoa, että edellä mainitut oletukset 1–3 ovat voimassa, vaikkei perusjoukkojen hajontoja tunneta, koska otoksista lasketut keskihajonnat ovat lähes samat.

Hypoteesit ovat seuraavat: $H_0 : \mu_1 = \mu_2$; $H_1 : \mu_2 > \mu_1$

$$s^2 = \frac{(22-1)\cdot 13{,}7^2 + (28-1)\cdot 13{,}9^2}{(22-1)+(28-1)} = 190{,}8 \text{ eli}$$

$$s = \sqrt{190{,}8} = 13{,}8$$

$$t = \frac{176{,}2 - 174{,}3}{13{,}8 \cdot \sqrt{\left(\frac{1}{22} + \frac{1}{28}\right)}} = 0{,}483$$

Johtopäätökset: Vapausasteiden lukumäärä on 22 + 28 – 2 = 48. Yksisuuntaisessa testissä *t*-jakauman kriittinen arvo on tällöin 2,403. Koska 0,483 < 2,403, nollahypoteesia ei hylätä.

Tulkinta: Tutkimuksen perusteella ei voida osoittaa, että länsisuomalaiset miehet olisivat keskimäärin pidempiä kuin itäsuomalaiset miehet.

ESIMERKKI 19

Haluttiin tutkia, eroavatko lukiota käyvien tyttöjen ja poikien matemaattiset taidot toisistaan. Asian selvittämiseksi järjestettiin matemaattisia taitoja mittaava koe, johon valittiin satunnaisesti 25 lukiolaista, joista 14 oli tyttöjä ja 11 poikia. Kokeen pistemäärät olivat seuraavat:

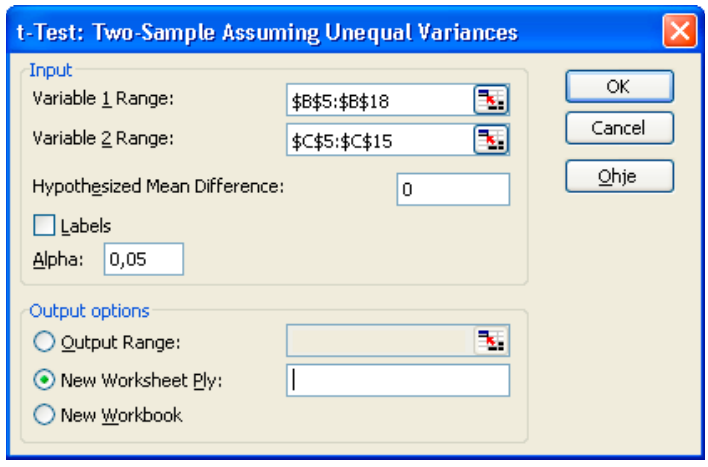
Tytöt: 36, 27, 19, 48, 29, 30, 18, 46, 37, 32, 22, 25, 40, 38

Pojat: 21, 11, 56, 47, 41, 19, 38, 35, 27, 53, 13.

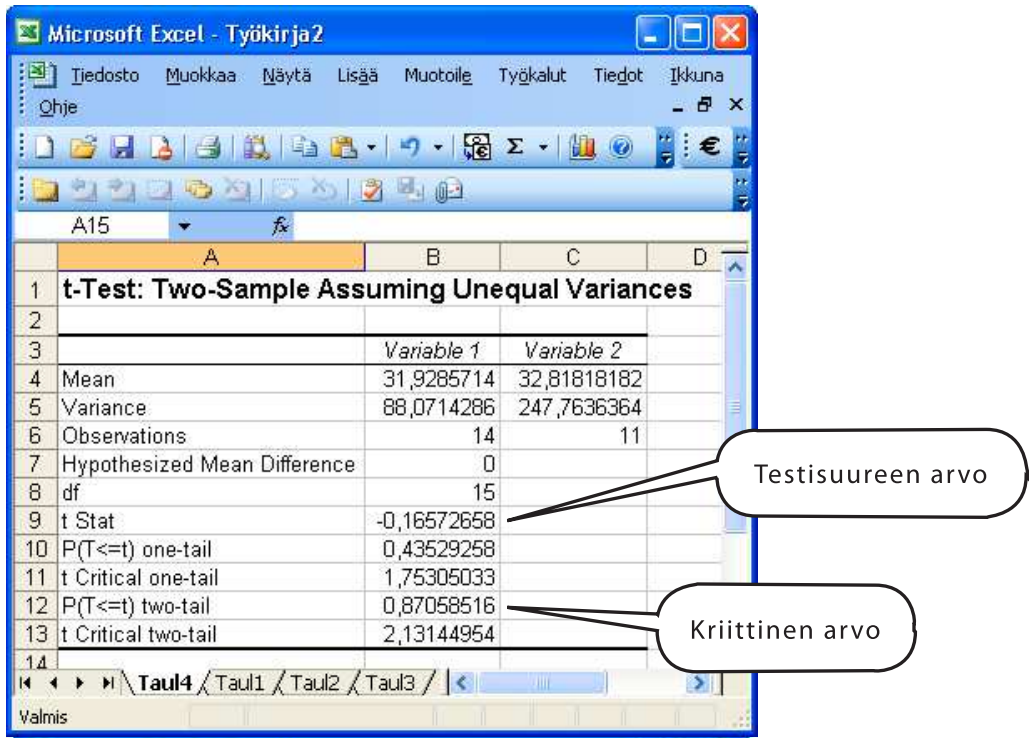
Testataan 5 %:n merkitsevyystasolla hypoteesi: Lukiota käyvien poikien ja tyttöjen matemaattiset taidot eivät poikkea toisistaan, kun oletetaan, että poikien ja tyttöjen matemaattiset taidot ovat normaalisti jakautuneita.

Merkitään tyttöjen kokeesta saamien pistemäärien keskiarvoa $\bar{x}_1$:llä ja poikien saamien pistemäärien keskiarvoa $\bar{x}_2$:lla. Tällöin hypoteesit ovat $H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 \neq \mu_2$.

Jo silmämääräisesti voidaan havaita, että poikien saamien pistemäärien vaihtelu on suurempi kuin tyttöjen pistemäärien vaihtelu, joten käytetään testiä, jossa keskihajonnat perusjoukoissa oletetaan erisuuriksi. Käytetään tehtävän ratkaisemiseen Excel-taulukkolaskentaa:



Tyttöjen saamat pistemäärät ovat Excel-taulukossa soluissa B5:B18 ja poikien pistemäärät soluissa C5:C15. Tulokseksi saadaan seuraava taulukko:



Johtopäätökset: Testisuureen arvo on –0,166, kun kriittinen arvo 5 %:n merkitsevyystasolla kaksisuuntaisessa testissä on –2,131. Nollahypoteesia ei siis voida hylätä.

Tulkinta: Tämän testin perusteella ei voida päätellä, että lukiota käyvien tyttöjen ja poikien keskimääräisissä matemaattisissa taidoissa olisi eroja.

Edellä olevissa tapauksissa otokset oletettiin keskenään riippumattomiksi. Edellisessä esimerkissä poikien testipistemäärät eivät millään tavalla vaikuta tyttöjen saamiin testipistemääriin tai päinvastoin.

Niin sanotussa klassisessa koeasetelmassa tutkitaan samaa koehenkilöiden ryhmää ennen tiettyä tutkimusta ja sen jälkeen. Tällöin havaintoarvot eivät ole toisistaan riippumattomia.

ESIMERKKI 20

Valmistaja väittää, että heidän valmistamansa kasvirasva pienentää veren kolesterolipitoisuutta. Väitteen oikeellisuuden toteamiseksi poimittiin satunnaisesti 10 koehenkilöä. Heidän kolesteroliarvonsa mitattiin, minkä jälkeen he korvasivat aikaisemmin käyttämänsä rasvan kyseisellä kasvirasvalla. Kymmenen viikon kuluttua mittaus tehtiin uudestaan. Tulokset olivat seuraavat:

Alussa	5,2	4,8	6,3	6,7	5,3	4,4	4,9	6,7	6,0	5,5
Kymmenen viikon kuluttua	4,5	4,8	5,7	5,8	4,8	3,7	4,3	6,6	5,3	5,0

Voidaanko luotettavasti päätellä, että kasvirasva pienentää veren kolesterolipitoisuutta?

Tutkimalla arvojen erotuksia havaitaan, että lähes poikkeuksetta kolesteroliarvot ovat pienentyneet. Näin näyttää siltä, että tuote saattaa pienentää veren kolesterolipitoisuutta. Jos erotus ei poikkea nollasta tai poikkeama on vähäinen, niin ei ole näyttöä, että muutosta on tapahtunut.

Verrannollisten parien t-testi

Poimitaan n yksikön otos perusjoukosta siten, että tehdään samoille tilastoyksiköille kaksi mittausta. Näin saadaan n toisistaan riippuvaa havaintoparia. Asetetaan nollahypoteesi H_0: Molempien mittaustulosten keskiarvo perusjoukossa on sama eli $\mu_1 = \mu_2$.

Oletetaan, että havaintoparien erotuksen jakauma on likimäärin normaalinen. Tällöin testisuure

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}, \text{ jossa}$$

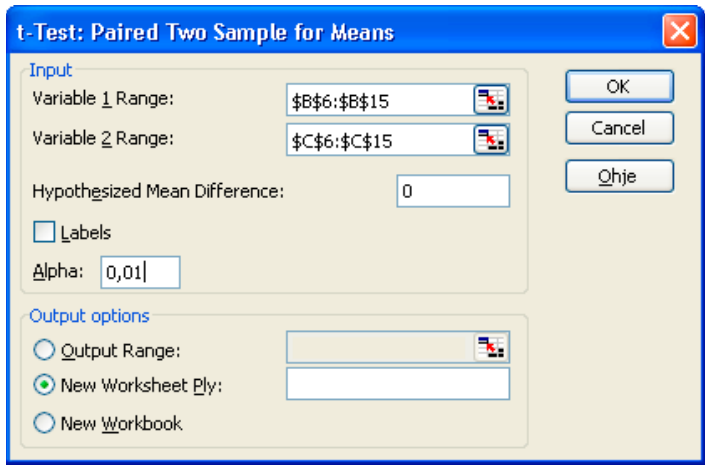
$\bar{d}$ = havaintoparien erotusten keskiarvo
s_d = erotusten keskihajonta
n = havaintoparien lukumäärä,

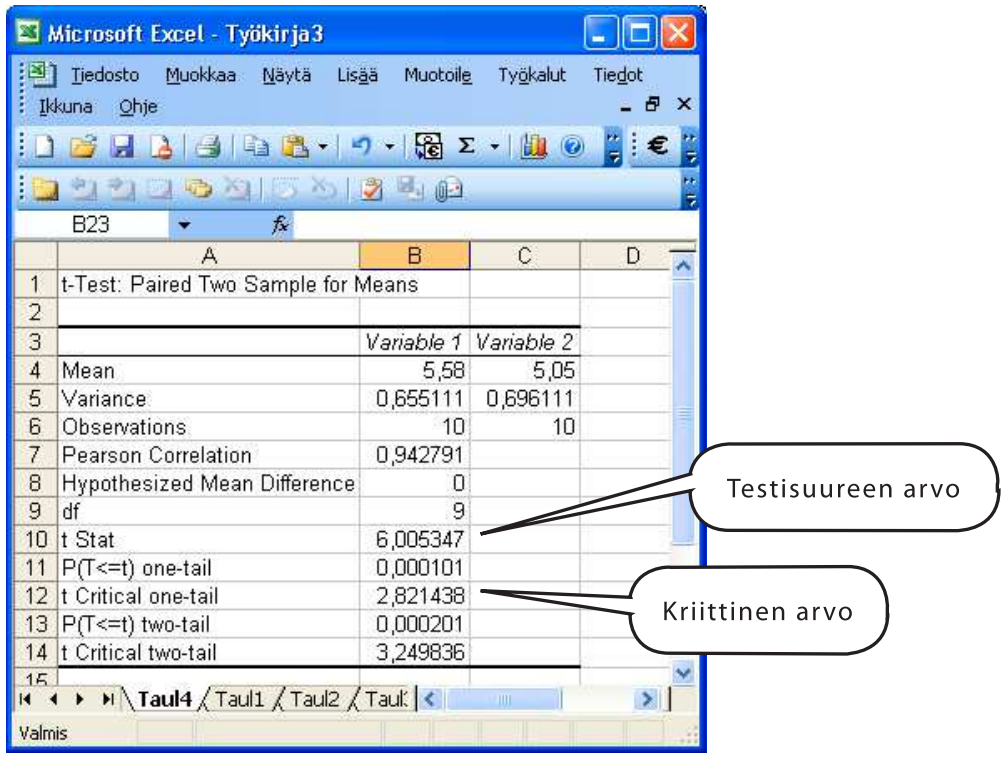
noudattaa t-jakaumaa vapausastein $n-1$, jos nollahypoteesi on tosi.

ESIMERKKI 21

Testaa yksisuuntaista testiä käyttämällä edellisen esimerkin hypoteesi H_0: Kasvirasva ei pienennä veren kolesterolipitoisuutta.

Ratkaistaan tehtävä Excelin avulla käyttämällä 1 %:n merkitsevyystasoa. Nollahypoteesi ja vaihtoehtoinen hypoteesi ovat $H_0 : \mu_1 = \mu_2; H_1 : \mu_2 < \mu_1$.





Johtopäätökset: Testisuureen arvo on 6,01. Kriittinen arvo on 2,82, joten nollahypoteesi hylätään 1 %:n merkitsevyystasolla. Ohjelman tulostama p-arvo 0,0001 osoittaa, että nollahypoteesia hylättäessä erehtymisriski on vain 0,01 %, joten tulos on tilastollisesti erittäin merkitsevä.

Tulkinta: Valmistajan ilmoitusta, että sen markkinoima kasvirasva pienentää veren kolesterolipitoisuutta, voidaan pitää luotettavana.

Monissa tutkimuksissa ollaan kiinnostuneita **suhteellisista osuuksista** ja niissä tapahtuneista muutoksista. Markkinatutkimuksissa selvitetään tuotteiden markkinaosuuksia, lähes kuukausittain julkaistaan tutkimuksia poliittisten puolueiden kannatuksen muutoksista jne.

Suhteellisen osuuden testauksessa käytettävä testimuuttuja liittyy toistokokeeseen A, jolla on kaksi tulosmahdollisuutta: 1) esiintyy, 2) ei esiinny. Kokeeseen liittyvä satunnaismuuttuja X noudattaa tällöin jakaumaa Bin(n, p). Muodostetaan uusi satunnaismuuttuja $P = X/n$, joka tarkoittaa tapahtuman A suhteellista osuutta n toistokokeessa.

Sivulla 144 esitetyn mukaan suurilla n:n arvoilla P:n jakauma on likimäärin $\mathrm{N}\left(p, \sqrt{\frac{p(1-p)}{n}} \right)$. Tällöin normitettu satunnaismuuttuja $Z = \frac{P - p}{\sqrt{\frac{p(1-p)}{n}}}$ noudattaa jakaumaa N(0, 1).

Suhteellisen osuuden testaus

Asetetaan hypoteesi H_0: Kun toistokoe tehdään n kertaa, tapahtuman A suhteellinen osuus on p_0.

Testimuuttuja $Z = \frac{P - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

noudattaa likimäärin jakaumaa N(0, 1), jos nollahypoteesi on tosi ja n on riittävän suuri.

ESIMERKKI 22

Tutkittiin, kuinka moni paikkakunnan asukas harrastaa liikuntaa säännöllisesti (vähintään kaksi kertaa viikossa). Tutkimusta varten haastateltiin 240 kuntalaista. Heistä 85 ilmoitti harrastavansa liikuntaa säännöllisesti. Voidaanko tutkimuksen perusteella päätellä, että yli 30 % kaikista kuntalaisista harrastaa säännöllistä liikuntaa?

Koska $n = 240$ ja $X = 85$, $P = \frac{85}{240} \approx 0{,}354$.

Käytetään yksisuuntaista testiä. Asetetaan hypoteesit

$H_0 : p_0 = 0{,}30; H_1 : p_0 > 0{,}30.$

$$Z = \frac{0{,}354 - 0{,}30}{\sqrt{\frac{0{,}30 \cdot (1 - 0{,}30)}{240}}} = 1{,}83$$

Johtopäätökset: Koska 1,83 > 1,64, joka on normaalijakaumaan liittyvä kriittinen arvo yksisuuntaisessa testissä 5 %:n merkitsevyystasolla, hylätään nollahypoteesi. Ero on tilastollisesti melkein merkitsevä.

Tulkinta: Voidaan päätellä, että kuntalaisista yli 30 % liikkuu säännöllisesti.

Tutkittaessa suhteellisen osuuden muutoksia kahden riippumattoman otoksen avulla käytetään **kahden otoksen testiä**.

Suhteellisen osuuden testaus kahdesta otoksesta

Merkitään tapahtuman A, joka voi esiintyä tai olla esiintymättä, suhteellisia osuuksia eri perusjoukoissa p_1 ja p_2.

Poimitaan perusjoukoista kaksi riippumatonta otosta. Ensimmäisen otoksen koko olkoon n_1 tilastoyksikköä ja toisen otoksen n_2 tilastoyksikköä. Ensimmäisessä otoksessa tapahtuman A suhteellinen osuus on P_1 ja toisessa otoksessa P_2.

Merkitään: $P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$.

Asetetaan hypoteesi H_0: Suhteelliset osuudet perusjoukoissa ovat yhtä suuret, eli $p_1 = p_2$.

Testimuuttuja $Z = \frac{P_1 - P_2}{\sqrt{P(1 - P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

noudattaa riittävän suurilla n_1:n ja n_2:n arvoilla likimäärin jakaumaa N(0, 1), kun nollahypoteesi on tosi.

ESIMERKKI 23

Tutkimuksessa haluttiin selvittää naisten ja miesten tupakoinnin eroja. Tutkimusta varten haastateltiin 342 miestä ja 417 naista. Miehistä tupakoitsijoita oli 166 ja naisista 187. Testaa 1 %:n merkitsevyystasolla hypoteesi, jonka mukaan tupakoivien osuus molemmissa ryhmissä on sama.

Lasketaan otosten perusteella tupakoivien osuus miesten ja naisten ryhmissä:

$$P_1 = \frac{166}{342} = 0,485;\ P_2 = \frac{187}{417} = 0,448.$$

Koska otosten perusteella näyttää siltä, että miesten tupakointi on yleisempää kuin naisten, asetetaan hypoteesit

$H_0: p_1 = p_2$; $H_1: p_1 > p_2$.

$$P = \frac{342 \cdot 0,485 + 417 \cdot 0,448}{342 + 417} = 0,465$$

$$Z = \frac{0,485 - 0,448}{\sqrt{0,465 \cdot (1 - 0,465) \cdot \left(\frac{1}{342} + \frac{1}{417}\right)}} = 1,02$$

Johtopäätökset: Koska 1,02 < 2,33, joka on normaalijakaumaan liittyvä kriittinen arvo yksisuuntaisessa testissä 1 %:n merkitsevyystasolla, nollahypoteesi jää voimaan. Nollahypoteesia ei voida hylätä myöskään 5 %:n merkitsevyystasolla (kriittinen arvo 1,64).

Tulkinta: Tutkimuksen perusteella ei voida päätellä, että miesten tupakointi olisi yleisempää kuin naisten tupakointi.

ESIMERKKI 24

Ratkaistaan edellinen esimerkki myös tilasto-ohjelmalla:

Hypothesis test for two independent proportions			
p_1	p_2	pc	
0,4854	0,4484	0,4651	p (as decimal)
166/342	187/417	353/759	p (as fraction)
166	187	353	X
342	417	759	n

0,0369	difference
0,0364	std. error
1,02	z
,1550	p-value (one-tailed)

Tulosteesta käy ilmi, että *p*-arvo on 0,155. Jos nollahypoteesi hylättäisiin, erehtymisriski olisi 15,5 %.

7.4 Jakaumasta riippumattomat testit

Edellisissä esimerkeissä tarkasteltiin **parametrisia** testejä. Niiden avulla testattiin hypoteeseja, jotka liittyivät perusjoukon parametreihin, kuten keskiarvoon tai prosenttiosuuteen. Läheskään aina parametrisia testejä ei kuitenkaan voida käyttää, koska tutkittavan muuttujan jakaumaa perusjoukossa ei tunneta (normaalijakaumaolettamus ei ole voimassa) tai muuttuja on mitattu liian karkealla mitta-asteikolla.

Jakaumasta riippumattomia testejä käytetään muun muassa jakauman muodon ja sijainnin selvittämiseen. Kun parametrisia testejä käytettäessä edellytyksenä oli, että tutkittavat muuttujat oli mitattu välimatka- tai suhdeasteikolla, jakaumasta riippumattomia testejä käytetään usein tutkittaessa **luokittelu**- ja **järjestysasteikon** muuttujia.

ESIMERKKI 25

Helsingin Sanomissa 22.8.2007 vertailtiin eri liikkeiden edullisuutta valitsemalla satunnaisesti 15 tuotetta. Näiden tuotteiden hinnat olivat helsinkiläisissä Lidl- ja Alepa-myymälöissä seuraavat:

Lidl	Alepa
0,19	0,30
0,56	0,59
0,66	0,75
0,59	0,54
0,63	0,79
0,78	0,90
1,55	1,59
0,99	0,99
1,99	2,10
0,99	1,60
1,09	1,29
1,69	2,39
0,79	2,00
2,39	2,69
0,59	0,89

Voidaanko päätellä, että jommankumman liikkeen hintataso on toista alhaisempi?

Yleinen tapa, jota monissa yhteyksissä käytetään, on laskea yhteen kaikkien tutkittavien tuotteiden hinnat (edellä 15,48 € ja 19,41 €). Tämä tapa ei kuitenkaan anna välttämättä oikeaa kuvaa hintatasosta. Jos esimerkiksi jossakin liikkeessä muutamien kalliiden tuotteiden hinnat ovat 10 % kalliimpia ja monet halpojen tavaroiden hinnat puolestaan 10 % halvempia, hintojen laskeminen yhteen voi antaa väärän kuvan liikkeen yleisestä hintatasosta kilpailijoihin verrattuna. Tilastollisesti tällaisissa tapauksissa on kysymys siitä, että liikkeissä myytävien tuotteiden hinta ei ole normaalisti jakautunut.

Parempi tapa tehdä hintavertailu on tutkia, sijaitsevatko hintojen jakaumat samassa kohdassa eli ovatko jakaumien **mediaanit** yhtä suuret.

Mann–Whitneyn testi

Oletetaan, että

1. satunnaismuuttujien X ja Y arvot $x_1, x_2,\ldots, x_{n1}$ ja $y_1, y_2,\ldots, y_{n2}$ ovat riippumattomia otoksia kahdesta perusjoukosta
2. muuttujat on mitattu vähintään **järjestysasteikolla**.

Asetetaan nollahypoteesi H_0: $Md_X = Md_Y$. Vaihtoehtoinen hypoteesi voi olla, että mediaanit ovat erisuuret tai toisen jakauman mediaani on suurempi kuin toisen jakauman mediaani.

Testaukseen käytetään Mann–Whitneyn testiä. Testisuureen jakauma riippuu otosten n_1 ja n_2 koosta siten, että suurilla otoksilla $(n_1, n_2 \geq 30)$ testisuure on likimäärin normaalisti jakautunut. Jos otosten koko on pieni $(n_1, n_2 < 30)$, testi perustuu havaintoarvojen järjestykseen.

Mann–Whitneyn testiä käytetään usein riippumattomien otosten *t*-testin sijasta, kun *t*-testin edellyttämä muuttujien normaalijakaumaoletus ei ole voimassa. Käytännössä testi tehdään aina tietokoneohjelman avulla.

ESIMERKKI 26

Testataan Mann–Whitneyn testin avulla, voidaanko edellisen esimerkin hintavertailun perusteella päätellä, että liikkeiden hintataso poikkeaa toisistaan. Käytetään yksisuuntaista testiä, koska näyttää siltä, että Lidlin hinnat olisivat hieman halvemmat kuin Alepan hinnat.

Asetetaan hypoteesit H_0: $Md_L = Md_A$; H_1: $Md_L < Md_A$, jossa

Md_L = Lidlin hintojen mediaani

Md_A = Alepan hintojen mediaani.

Mann–Whitney Test

n	sum of ranks	
15	209	Lidl
15	256	Alepa
30	465	total
	232,50	expected value
	24,08	standard deviation
	-0,98	z, corrected for ties
	,1646	p-value (one-tailed)

Johtopäätökset: Testimuuttujan arvo on –0,98 ja *p*-arvo 0,1646. Jos nollahypoteesi hylättäisiin, erehtymisriski olisi 16 %.

Tulkinta: Testin perusteella ei voida päätellä, että Lidlin ja Alepan hintatasot poikkeaisivat toisistaan, joskin hieman sen suuntaisia viitteitä on. Erehtymisriski on 16 %, jos päätellään, että Lidl on halvempi. Asian tutkimista voitaisiin jatkaa kasvattamalla otoksen kokoa.

Aikaisemmin käsiteltiin verrannollisten parien *t*-testiä. Jotta sitä voitaisiin käyttää, verrannollisten parien erotuksen on noudatettava normaalijakaumaa. Ellei tämä vaatimus toteudu, voidaan verrannollisten parien *t*-testin asemasta käyttää Wilcoxonin testiä.

Wilcoxonin testi

Poimitaan n yksikön otos perusjoukosta siten, että tilastoyksiköiden arvot mitataan kahteen kertaan (ennen–jälkeen). Mitta-asteikkona on vähintään **järjestysasteikko**. Asetetaan nollahypoteesi H_0: Ryhmät eivät eroa sijainniltaan toisistaan eli muutosta ei ole tapahtunut.

Testaukseen käytetään Wilcoxonin testiä. Testisuure lasketaan vastinparien erotusten avulla siten, että aluksi erotukset pannaan itseisarvojen suuruuden perusteella järjestykseen ja näille erotuksille annetaan sen perusteella järjestysluvut, minkä jälkeen järjestysluvuille annetaan erotuksen etumerkki. Sellaisia havaintoja, joiden erotukset ovat nollia, ei oteta mukaan. Testisuure riippuu otoksen koosta siten, että suurilla otoksilla ($n \geq 30$) testisuureen jakauma on likimäärin N(0, 1), jos nollahypoteesi on tosi.

Käytännössä tämäkin testi tehdään aina tietokoneohjelmalla.

ESIMERKKI 27

Tarkastellaan uudestaan esimerkkiä 20, jossa tutkittiin, pienentääkö kasvirasva veren kolesterolipitoisuutta. Koehenkilöiden kolesterolia arvot mitattiin, minkä jälkeen he korvasivat aikaisemmin käyttämänsä rasvan kasvirasvalla. Kymmenen viikon kuluttua mittaus tehtiin uudestaan. Tulokseksi saatiin seuraavat havaintoarvot:

Alussa	5,2	4,8	6,3	6,7	5,3	4,4	4,9	6,7	6,0	5,5
Kymmenen viikon kuluttua	4,5	4,8	5,7	5,8	4,8	3,7	4,3	6,6	5,3	5,0

Kun tutkitaan muuttujien arvojen erotusta ja sen jakaumaa, ei voida olla varmoja, että muuttujien erotus noudattaa normaalijakaumaa, joten on varminta käyttää Wilcoxonin testiä.

Asetetaan nollahypoteesi H_0: Kasvirasva ei pienennä veren kolesterolipitoisuutta, eli muuttujien jakaumien sijainti on sama. Vaihtoehtoinen hypoteesi on, että kyseinen rasva pienentää veren kolesterolipitoisuutta, joten testi on yksisuuntainen.

Wilcoxon Signed Rank Test	
variables:	Aluksi – Myöhemmin
45	sum of positive ranks
0	sum of negative ranks
9	n
22,50	expected value
7,50	standard deviation
3,00	z, corrected for ties
,0013	p-value (one-tailed)

Johtopäätökset: Tulosteesta käy ilmi, että testimuuttujan arvo on 3,00 ja *p*-arvo 0,0013. Jos nollahypoteesi hylätään, erehtymisriski on siten 0,13 %. Mittaustulosten perusteella kolesterolipitoisuuksien ero on tilastollisesti merkitsevä.

Tulkinta: Wilcoxonin testi tukee myös valmistajan väitettä, että sen markkinoima kasvirasva pienentää veren kolesterolipitoisuutta.

Yksi eniten käytetyistä jakaumasta riippumattomista testeistä on χ^2-testi. χ^2-yhteensopivuustestin avulla tutkitaan, ovatko satunnaisilmiöön liittyvät havaintoarvot sopusoinnussa teoreettisen mallin kanssa.

χ^2-yhteensopivuustesti

Oletetaan, että tapahtumiin $E_1, E_2,..., E_k$ liittyvät havaitut frekvenssit o_1, $o_2,..., o_k$ ja teoreettiset frekvenssit $e_1, e_2,..., e_k$. Testimuuttuja

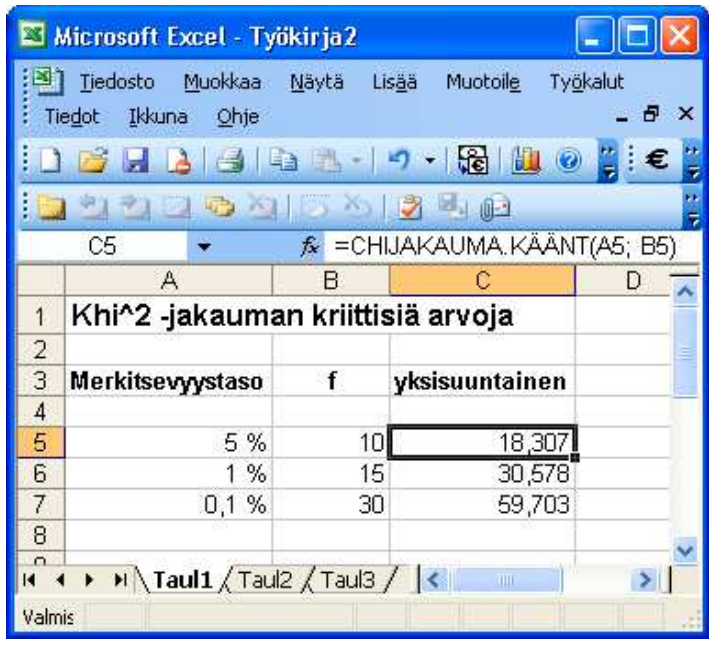
$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + ... + \frac{(o_k - e_k)^2}{e_k} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

noudattaa χ^2-jakaumaa vapausastein $f = k - 1$, jos nollahypoteesi on tosi.

Jotta testimuuttuja noudattaisi χ^2-jakaumaa, pitäisi **kaikkien teoreettisten frekvenssien $e_i > 5$**. Jos jotkin teoreettiset frekvenssit jäävät tämän rajan alapuolelle, testi voidaan usein tehdä, kun tällaisia frekvenssejä on **enintään 20 % kaikista teoreettisista frekvensseistä**. Kaikkien teoreettisten frekvenssien on kuitenkin oltava **suurempia kuin 1**. Mitä suurempia teoreettiset frekvenssit ovat, sitä paremmin testimuuttuja noudattaa χ^2-jakaumaa. Myös χ^2-jakauman kriittisille arvoille on taulukko kirjan liitteenä (ks. liitettä χ^2-jakaumaan liittyvät kriittiset arvot).

χ^2-jakauman kriittiset arvot voidaan määrittää myös Excelin avulla. Excelin funktio CHIJAKAUMA.KÄÄNT(todennäköisyys; vapausasteet) (engl. CHIINV) laskee χ^2-jakaumaan liittyvät kriittiset arvot yksisuuntaisessa testissä.

Seuraavaan taulukkoon on laskettu eräitä χ^2-jakauman kriittisiä arvoja:



ESIMERKKI 28

Paikkakunnan terveyskeskuksella on neljä terveysasemaa. Tutkimuksessa haluttiin selvittää, onko henkilökunnan kuormitus kaikilla terveysasemilla sama. Asian selvittämiseksi tilastoitiin satunnaisesti valitun kuukauden kävijät kullakin asemalla.

Terveysasema	A	B	C	D
Kävijämäärä (henk./kk)	1 189	1 104	1 156	1 063

Testaa 5 %:n merkitsevyystasolla hypoteesi H_0: Terveysasemien kävijämäärät eivät poikkea toisistaan.

Tutkimuskuukauden aikana terveysasemilla kävi yhteensä 1 189 + 1 104 + 1 156 + 1 063 = 4 512 henkilöä. Jos kaikilla terveysasemilla olisi yhtä paljon kävijöitä, jokaisen terveysaseman kävijämäärän pitäisi olla $\frac{4\,512}{4} = 1128$, eli $e_i = 1\,128$, kun $i = 1, 2, 3$ tai 4.

Testin avulla selvitetään, johtuvatko kävijämäärien erot tutkimukseen valitun kuukauden aikana satunnaisista tekijöistä vai voidaanko niiden perusteella päätellä, että ero on pysyvä.

Testimuuttuja

$$\chi^2 = \frac{(1189-1128)^2}{1128} + \frac{(1104-1128)^2}{1128} + \frac{(1156-1128)^2}{1128} +$$

$$\frac{(1063-1128)^2}{1128} = 8{,}25.$$

Johtopäätökset: Taulukosta käy ilmi, että χ^2-jakauman kriittinen arvo 5 %:n merkitsevyystasolla vapausastein 4 – 1 = 3 on 7,815. Koska 8,25 > 7,815, nollahypoteesi hylätään.

Tulkinta: Tutkimuksen perusteella voidaan päätellä, että terveysasemien kävijämäärissä on eroja.

ESIMERKKI 29

Ratkaistaan edellinen esimerkki myös tietokoneohjelmalla:

Goodness of Fit Test

observed	expected	O – E	$(O - E)^2 / E$	% of chisq
1189	1128,000	61,000	3,299	39,98
1104	1128,000	–24,000	0,511	6,19
1156	1128,000	28,000	0,695	8,42
1063	1128,000	–65,000	3,746	45,40
4512	4512,000	0,000	8,250	100,00

8,25	chi-square
3	df
,0411	p-value

Tulosteen *p*-arvo 0,0411 osoittaa, että erehtymisriski nollahypoteesia hylättäessä on 4,1 %. Kävijämäärien ero on siten tilastollisesti melkein merkitsevä.

χ^2-riippumattomuustesti

χ^2-testiä käytetään myös **riippumattomuustestinä**, kun halutaan testata, riippuvatko kaksi tilastollista muuttujaa toisistaan. Mitta-asteikoksi riittää **luokitteluasteikko**. Testiä varten muodostetaan havaituista frekvensseistä ristiintaulukoimalla **kontingenssitaulukko**.

Testi tehdään χ^2-testimuuttujan avulla vapausastein $(m-1)(n-1)$, jossa m on taulukon vaakarivien ja n pystyrivien (sarakkeiden) lukumäärä.

Asetetaan nollahypoteesi H_0: Tilastolliset muuttujat ovat riippumattomia.

Testimuuttuja $\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$

noudattaa χ^2-jakaumaa vapausastein $(m-1)(n-1)$, jos nollahypoteesi on tosi.

ESIMERKKI 30

Tutkimuksella selvitettiin kansalaisten tyytyväisyyttä (1 = erittäin tyytyväinen, 2 = jokseenkin tyytyväinen, 3 = neutraali, 4 = jokseenkin tyytymätön ja 5 = erittäin tyytymätön) hallituksen politiikkaan.

Vastausten jakauma oli seuraava:

	1	2	3	4	5	Yhteensä
Mies	15	23	48	59	23	168
Nainen	6	14	62	98	42	222
Yhteensä	21	37	110	157	65	390

Testataan 5 %:n merkitsevyystasolla, riippuuko tyytyväisyys hallituksen politiikkaan sukupuolesta. Asetetaan hypoteesit:

H_0: Tyytyväisyys hallituksen politiikkaan on riippumaton sukupuolesta.

H_1: Tyytyväisyydellä hallituksen politiikkaan ja sukupuolella on riippuvuutta.

Taulukossa olevat luvut 15, 23, 48, 59, 23, 6, 14, 62, 98 ja 42 ovat havaittuja frekvenssejä o_i. Niitä on verrattava teoreettisiin frekvensseihin e_i, jotka saadaan kätevimmin, kun kerrotaan kutakin solua vastaavat rivi- ja sarakesummat keskenään ja jaetaan tulo kokonaisfrekvenssillä 390.

Teoreettiset frekvenssit ovat seuraavat:

	1	2	3	4	5
Mies	$\frac{168 \cdot 21}{390} = 9,0$	$\frac{168 \cdot 37}{390} = 15,9$	$\frac{168 \cdot 110}{390} = 47,4$	$\frac{168 \cdot 157}{390} = 67,6$	$\frac{168 \cdot 65}{390} = 28,0$
Nainen	$\frac{222 \cdot 21}{390} = 12,0$	$\frac{222 \cdot 37}{390} = 21,1$	$\frac{222 \cdot 110}{390} = 62,6$	$\frac{222 \cdot 157}{390} = 89,4$	$\frac{222 \cdot 65}{390} = 37,0$

Testimuuttujan arvo

$$\chi^2 = \frac{(15 - 9,0)^2}{9,0} + \frac{(23 - 15,9)^2}{15,9} + \ldots + \frac{(42 - 37,0)^2}{37,0} = 15,9.$$

Vapausasteiden lukumäärä $f = (2 - 1) \cdot (5 - 1) = 4$.

Johtopäätökset: χ^2-jakauman kriittinen arvo 5 %:n merkitsevyystasolla vapausastein 4 on 9,488. Koska 15,9 > 9,488, nollahypoteesi hylätään.

Tulkinta: Tämän tutkimuksen mukaan sukupuolella ja tyytyväisyydellä hallituksen politiikkaan on keskinäistä riippuvuutta. Tutkimalla teoreettisia ja havaittuja frekvenssejä huomataan, että naiset ovat keskimäärin tyytymättömämpiä hallitukseen kuin miehet.

ESIMERKKI 31

Edellinenkin tehtävä on luonnollisesti kätevintä ratkaista tietokoneohjelmalla:

Chi-square Contingency Table Test for Independence

		1	2	3	4	5	Total
Mies	Observed	**15**	**23**	**48**	**59**	**23**	168
	Expected	9,05	15,94	47,38	67,63	28,00	168,00
Nainen	Observed	**6**	**14**	**62**	**98**	**42**	222
	Expected	11,95	21,06	62,62	89,37	37,00	222,00
Total	Observed	21	37	110	157	65	390
	Expected	21,00	37,00	110,00	157,00	65,00	390,00

15,90	chi-square
4	df
,0032	p-value

Tulkinta: Tulosteen *p*-arvo 0,0032 osoittaa, että erehtymisriski on vain 0,3 %, jos nollahypoteesi hylätään. Ero miesten ja naisten mielipiteissä on siten tilastollisesti merkitsevä. Naiset suhtautuvat tämän tutkimuksen perusteella hallituksen politiikkaan huomattavasti kriittisemmin kuin miehet.

7.5 Testin valinta

Tilastollinen testaus käsittää tiivistetysti seuraavat vaiheet:

1. hypoteesin asettaminen
2. otoksen poimiminen
3. tunnuslukujen tai havaittujen frekvenssien laskeminen otoksesta
4. testin valinta
5. testimuuttujan arvon laskeminen
6. testimuuttujan arvon vertaaminen todennäköisyysjakauman kriittisiin arvoihin
7. tuloksen tulkinta ja johtopäätösten tekeminen.

Varsinaisessa testauksessa ratkaisevin vaihe on tapaukseen sopivan testin valinta. Testimuuttujan laskeminen on helppoa, mutta testin valintaan ja tulosten tulkintaan on kiinnitettävä erityistä huomiota. Käytettävä testi voidaan valita monella eri periaatteella. Edellä mainittiin jo jako parametrisiin ja jakaumasta riippumattomiin testeihin.

Parametrisia testejä käytetään, kun testataan perusjoukon jonkin parametrin (keskiarvo, keskihajonta, suhteellinen osuus) arvoa. Tällöin on kiinnitettävä huomiota **otoksesta lasketun tunnusluvun jakaumaan**, joka riippuu muun muassa **otoksen koosta**. Pienissä otoksissa jakauma on usein ***t*-jakauma,** ellei perusjoukon keskihajontaa tunneta. Erityistä huomiota on tällöin kuitenkin kiinnitettävä *t*-testin edellyttämään perusjoukon normaalijakaumaoletukseen. Ellei voida olla varmoja, että satunnaismuuttuja perusjoukossa on normaalisti jakautunut, kannattaa *t*-testin sijasta käyttää tapauksesta riippuen joko Mann–Whitneyn testiä tai Wilcoxonin testiä.

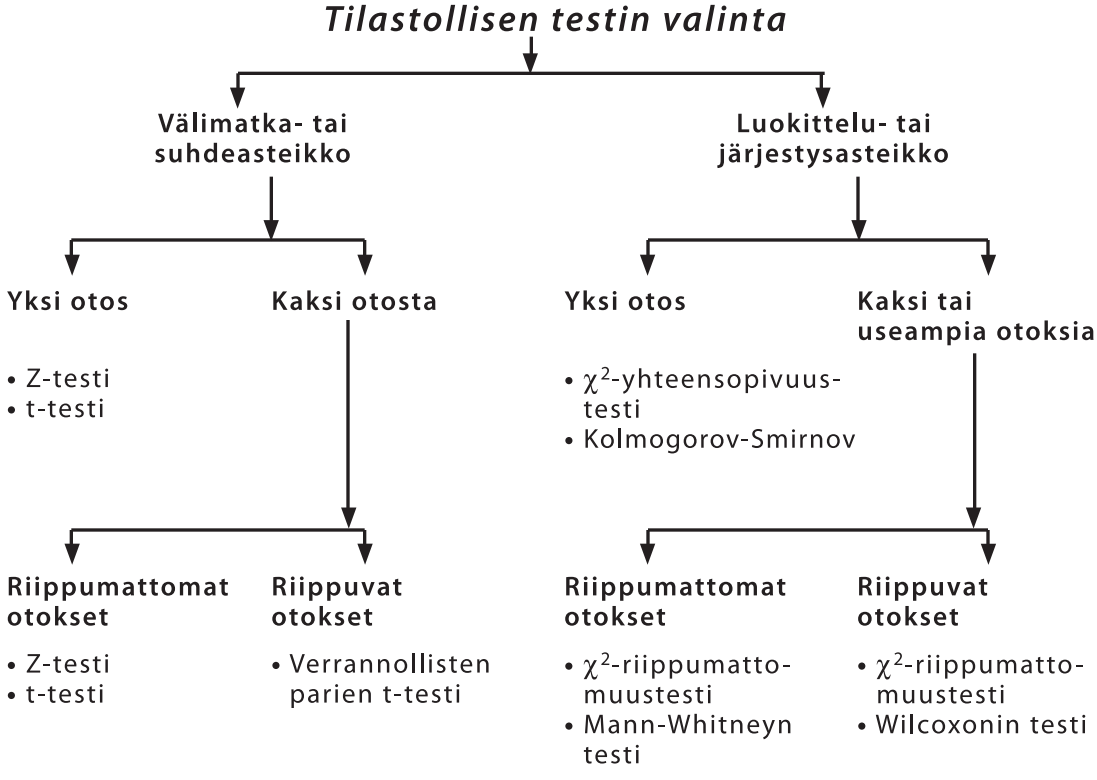
Perusjoukon normaalisuutta voi tutkia yksinkertaisimmillaan piirtämällä otosjakaumasta histogrammin ja vertaamalla sitä Gaussin käyrään. Useisiin tilasto-ohjelmiin (mm. SPSS) sisältyy Kolmogorov–Smirnovin testi, jonka avulla perusjoukon normaalijakaumaoletusta voi kätevästi testata. Suurista otoksista laskettujen tunnuslukujen jakauma on keskeisen raja-arvolauseen perusteella likimäärin **normaalijakauma**. Jotta parametrisia testejä voidaan käyttää, tutkittavan muuttujan pitää olla mitattu **välimatka**- tai **suhdeasteikolla**.

Jakaumasta riippumattomista testeistä χ^2-testiä käytetään, kun halutaan selvittää muuttujan jakaumaa havaittujen frekvenssien avulla tai **luokittelu**- tai **järjestysasteikolla** mitattujen muuttujien riippumattomuutta. Kuten edellä mainittiin, Mann–Whitneyn testiä tai Wilcoxonin testiä käytetään usein silloin, kun *t*-testin käytön edellytykset eivät ole voimassa.

Jos on mahdollista valita usean testin joukosta, valinnan ratkaisee testin **voimakkuus**, jolla tarkoitetaan sen kykyä hylätä väärä nollahypoteesi. Yleensä parametriset testit ovat voimakkaampia kuin jakaumasta riippumattomat testit. Kannattaakin esimerkiksi käyttää mieluummin verrannollisten parien *t*-testiä kuin Wilcoxonin testiä, jos *t*-testin käytön edellytykset ovat voimassa.

Koska tilastolliset testit perustuvat aina otokseen, johtopäätökset eivät koskaan ole täysin varmoja. Testien yhteydessä arvioidaankin aina, kuinka suuri **riski** on**, että otoksen perusteella on tehty perusjoukosta virheellinen johtopäätös**. On myös muistettava, että tilastollisen päätöksenteon yhteydessä joudutaan usein tekemään oletuksia, joiden oikeellisuus ei matemaattisesti ole sataprosenttista. Tilastolliset testit ovat siten ainoastaan apuneuvo päätöksenteossa.

Testin valintaa voidaan kuvata seuraavalla puurakenteella.



7.6 Case A: Tavaratalon palvelutasotutkimus

Seuraavassa esitetään todelliseen tutkimukseen liittyvä case tutkimusraportteineen. Monet yritykset tekevät tai teettävät vastaavia tutkimuksia säännöllisesti.

Tutkimuksen tausta

Kaupungin keskustassa sijaitseva osuuskaupan omistama tavaratalo haluaa tehdä tutkimuksen asiakaspalvelustaan. **Tutkimuksen päätarkoitus on selvittää asiakaspalvelun nykyinen taso, koska liike kilpailee palvelun laadulla.** Liikkeen tavoitteena on olla parhaiten asiakkaita palveleva tavaratalo kaupungissa. Tutkimuksen tulosten perusteella ryhdytään tarpeellisiin toimiin tavoitteen saavuttamiseksi.

Tavaratalon hintataso on korkeampi kuin kaupungin laidoilla sijaitsevissa marketeissa, ja tavaratalossa on myös vähemmän erikoistarjouksia. Tutkittavilta asiakkailta tiedustellaan mielipiteitä näistäkin tekijöistä, vaikkeivät ne olekaan tavaratalon kilpailuvaltteja.

Lisäksi tutkimuksella selvitetään asiakaskunnan rakennetta. Tutkittavia asioita ovat asiakaskunnan sukupuoli- ja ikäjakaumat. Samoin tutkitaan, kuinka suuri osa asiakkaista on osuuskaupan jäseniä (vakioasiakkaita), mitkä ovat asiakaskunnan muodostavien kotitalouksien nettotulot kuukaudessa ja kuinka suuri osa asiakaskotitalouksista on lapsiperheitä.

Tutkimuksen toteuttaa kaupungissa toimiva markkinatutkimuksia tekevä yritys. Perustiedot kerätään **haastattelemalla** (ks. lomake s. 208). Otantamenetelmänä käytetään **systemaattista otantaa**, ja otoksen koko on **100 tilastoyksikköä**. Lopuksi tutkimuksesta laaditaan raportti toimenpidesuosituksineen.

Asiakaskysely

Haluamme palvella asiakkaitamme mahdollisimman hyvin ja toivomme, että kertoisit meille myymälämme asiakaspalvelusta. Mielipiteesi on meille tärkeä, ja kaikki palaute, niin positiivinen kuin negatiivinenkin, otetaan huomioon.

Kiitos vaivannäöstäsi!

Ympyröi sopiva vaihtoehto.

1 Vastaaja on
1 nainen
2 mies

2 Vastaajan ikä vuosina on
1 alle 20
2 20–29
3 30–39
4 40–49
5 50–59
6 60 tai enemmän

3 Vastaajan taloudessa
1 ei ole lapsia
2 on lapsia

4 Vastaajan kotitalouden nettotulot ovat (€/kk)
1 alle 800
2 800–1 599
3 1 600–2 399
4 2 400–3 199
5 3 200–3 999
6 4 000 tai enemmän

5 Vastaaja
1 on osuuskaupan jäsen
2 ei ole jäsen

Seuraaviin kysymyksiin vastatessasi ympyröi se vaihtoehto (vain yksi), joka lähinnä vastaa mielipidettäsi. Vaihtoehdot ovat **1 = täysin eri mieltä, 2 = jokseenkin eri mieltä, 3 = neutraali mielipide, 4 = jokseenkin samaa mieltä, 5 = täysin samaa mieltä**.

6	**Myymälä on viihtyisä**	1	2	3	4	5
7	**Tuotteet on helppo löytää**	1	2	3	4	5
8	**Hinnat ovat hyvin esillä**	1	2	3	4	5
9	**Henkilökunta on ystävällistä**	1	2	3	4	5
10	**Liikkeessä on hyvät vaihtuvat erikoistarjoukset**	1	2	3	4	5
11	**Liikkeen hintataso on edullinen**	1	2	3	4	5
12	**Liikkeessä asiointi on nopeaa**	1	2	3	4	5
13	**Liikkeen tuotevalikoima on hyvä**	1	2	3	4	5
14	**Liikkeen myyjien asiantuntemus on hyvä**	1	2	3	4	5
15	**Liikkeen aukioloajat ovat sopivat**	1	2	3	4	5

Haastattelulomakkeen laativat yhteistyönä tavaratalon ja tutkimuksen tekevän yrityksen edustajat. Liikkeen edustajat haluavat tutkimuksen avulla selvittää

– liikkeen viihtyisyyttä
– tuotteiden löydettävyyttä
– hintojen esillä oloa
– henkilökunnan ystävällisyyttä
– asioinnin nopeutta
– liikkeen tuotevalikoimaa
– myyjien asiantuntemusta
– aukioloaikojen sopivuutta.

Tulosten analysointi

Tutkimuksen perustiedoista on koottu havaintomatriisi **KYSELY**, jonka saat käyttöösi Sanoma Pron kotisivuilta osoitteesta www.sanomapro.fi > Oppilaitoksille > Ammatti- ja korkeakoulut >Ammattiverkko.

Raportissa on tulostettu taustamuuttujien 1–5 jakaumat niitä havainnollistavine kuvioineen ja muuttujien 6–15 jakaumien perusteella laadittava **palvelutasoprofiili**. Palvelutasoprofiili saadaan laskemalla muuttujien **keskiarvot**. Ahtaan tilastollisesti tulkittuna keskiarvojen sijasta tulisi käyttää **mediaaneja**, sillä muuttujat on mitattu järjestysasteikolla. Tällöin kuitenkin menetettäisiin informaatiota, koska muuttujien mediaanit ovat kokonaislukuja. Mainittakoon, että keskiarvojen käyttö järjestysasteikolla mitattujen muuttujien profiilien esittämisessä on yleinen tapa asiakaspalvelu- ja markkinointitutkimuksissa. Palvelutasoprofiili tulostetaan myös kuviona.

Muuttujat, jotka palvelutasoprofiilissa saavat lukua 3 pienemmän arvon, tutkitaan muita yksityiskohtaisemmin, sillä näissä palvelutekijöissä on runsaasti parantamisen varaa.

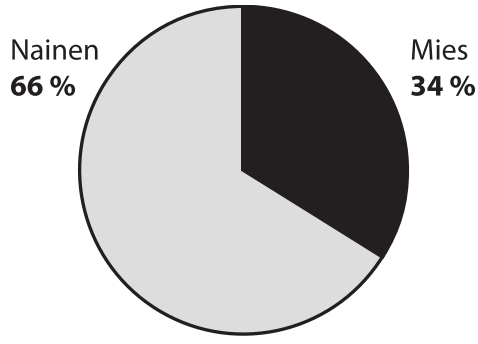
Asiakaskunnan rakenne

Kysymysten 1–5 avulla selvitettiin asiakaskunnan rakennetta. Vastausten jakaumaa tarkastellaan kysymyksittäin.

1 Vastaaja on
1 nainen
2 mies

Tutkimukseen osallistuneiden sukupuolijakauma on seuraava:

Sukupuoli	*f*	%
1	66	66,0
2	34	34,0
Yhteensä	100	100,0



Vastaajien jakauma sukupuolen mukaan.

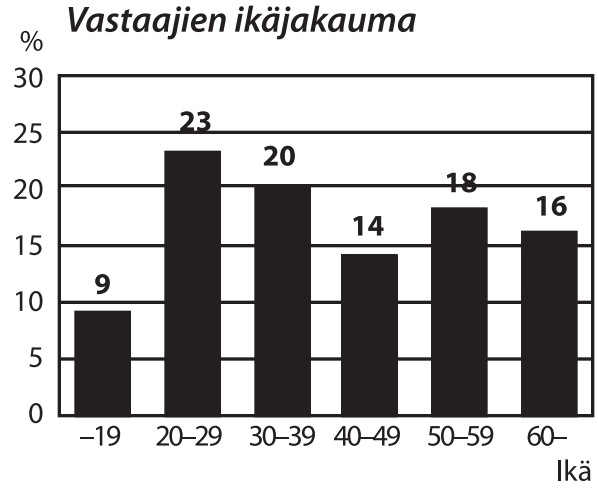
Asiakkaista noin kaksi kolmannesta on naisia ja yksi kolmannes miehiä. Koska naiset muodostavat tavaratalon pääasiakaskunnan, heidän toiveidensa huomioon ottaminen asiakaspalvelussa on ensisijaisen tärkeää.

2 Vastaajan ikä vuosina on

1 alle 20
2 20–29
3 30–39
4 40–49
5 50–59
6 60 tai enemmän

Vastanneiden ikäjakauma on seuraava:

Ikä (v)	*f*	%
alle 20	9	9,0
20–29	23	23,0
30–39	20	20,0
40–49	14	14,0
50–59	18	18,0
60 tai enemmän	16	16,0
Yhteensä	100	100,0



Tutkimukseen osallistuneiden iät jakautuivat melko tasaisesti kaikkiin ikäryhmiin, joten muun muassa liikkeen erikoistarjoukset pitäisi suunnata eri ikäryhmille. Myös tuotevalikoiman pitää olla laaja.

3 Vastaajan taloudessa

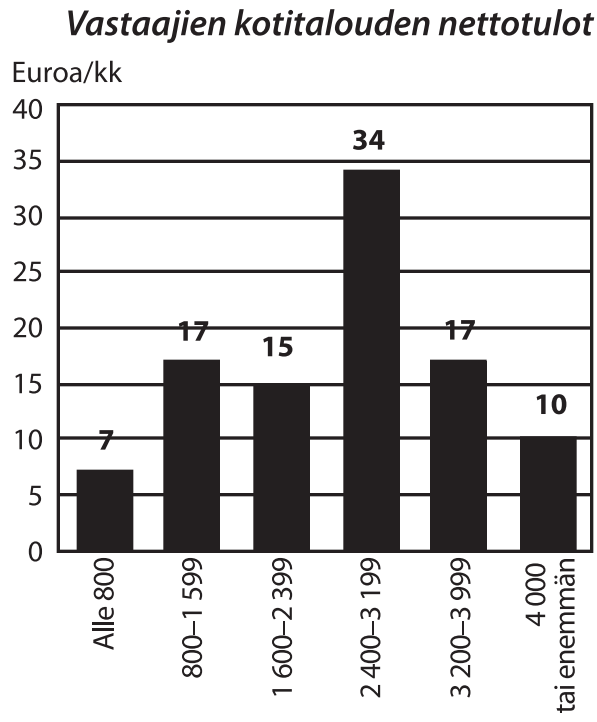
1 ei ole lapsia
2 on lapsia

Lapset	*f*	%
1	58	58,0
2	42	42,0
Yhteensä	100	100,0

Lapsettomia talouksia on asiakaskunnassa jonkin verran enemmän kuin lapsiperheitä. Myös lapsiperheet ovat kuitenkin merkittävä asiakasryhmä, jonka palveluun on kiinnitettävä huomiota.

4 Vastaajan kotitalouden nettotulot ovat (€/kk)

1 alle 800
2 800–1 599
3 1 600–2 399
4 2 400–3 199
5 3 200–3 999
6 4 000 tai enemmän



Nettotulot	*f*	%
alle 800	7	7,0
800–1 599	17	17,0
1 600–2 399	15	15,0
2 400–3 199	34	34,0
3 200–3 999	17	17,0
4 000 tai enemmän	10	10,0
	100	100,0

Hyvätuloisten osuus tavaratalon asiakaskunnassa on melko suuri. Vähintään 2 400 euron nettotulot kuukaudessa oli 61 prosentilla vastaajien kotitalouksista. Liikkeen tuotevalikoimassa on sen vuoksi oltava myös laadukkaita ja ylellisiä tuotteita, joissa hinta ei ole ratkaiseva tekijä.

5 Vastaaja

1 on osuuskaupan jäsen
2 ei ole jäsen

Jäsenyys	*f*	%
1	53	53,0
2	47	47,0
Yhteensä	100	100,0

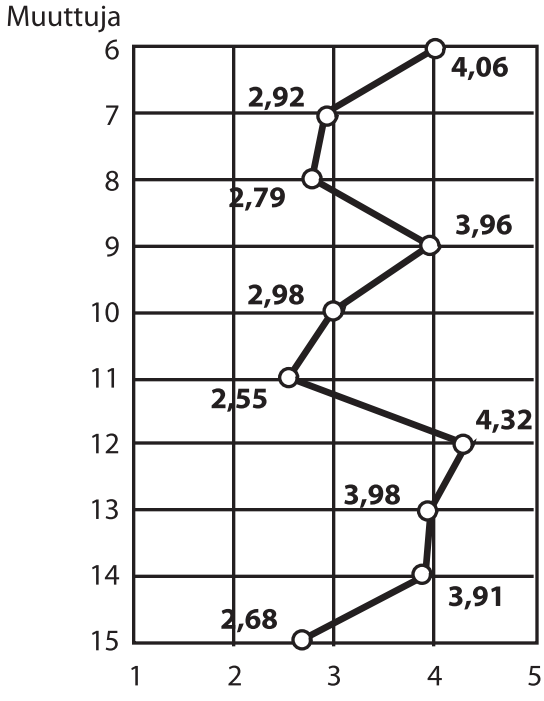
Osuuskaupan jäseniä (vakioasiakkaita) on yli puolet tutkimukseen osallistuneista.

Mielenkiintoinen tutkimuskohde on kysymys, poikkeavatko jäsenten ja ei-jäsenten mielipiteet asiakaspalvelun tasosta toisistaan.

Palvelutasoprofiili

Muuttujien 6–15 keskiarvot ovat seuraavat:

6	**Myymälä on viihtyisä**	4,06
7	**Tuotteet on helppo löytää**	2,92
8	**Hinnat ovat hyvin esillä**	2,79
9	**Henkilökunta on ystävällistä**	3,96
10	**Liikkeessä on hyvät vaihtuvat erikoistarjoukset**	2,98
11	**Liikkeen hintataso on edullinen**	2,55
12	**Liikkeessä asiointi on nopeaa**	4,32
13	**Liikkeen tuotevalikoima on hyvä**	3,98
14	**Liikkeen myyjien asiantuntemus on hyvä**	3,91
15	**Liikkeen aukioloajat ovat sopivat**	2,68



Palvelutasoprofiili.

Asiakkaat ovat keskimäärin hyvin tyytyväisiä **myymälän viihtyisyyteen, henkilökunnan ystävällisyyteen, liikkeessä asioinnin nopeuteen, tuotevalikoimaan ja myyjien asiantuntemukseen**.

Sen sijaan **tuotteiden löydettävyys, hintojen esillä olo, liikkeen erikoistarjoukset, hintataso ja aukioloajat** saivat keskiarvokseen alle 3,00. Niissä on siis ainakin jonkin verran korjaamisen varaa.

Jo aikaisemmin todettiin, ettei tavaratalo pyrikään kilpailemaan hinnoilla tai erikoistarjouksilla keskustan ulkopuolella sijaitsevien marketien kanssa. Muuttujat 10 ja 11 jätetään siten pois jatkotarkasteluista.

Ryhdytään tutkimaan muuttujia 7 (tuotteet on helppo löytää), 8 (hinnat ovat hyvin esillä) ja 15 (liikkeen aukioloajat ovat sopivat) yksityiskohtaisesti, jotta saataisiin selville, miksi niiden keskiarvot jäivät alle kolmen.

Tuotteiden löydettävyys

Voisi olettaa, että ainakin kanta-asiakkaat löytävät tuotteet vaivattomasti. Tutkitaan ristiintaulukoimalla ja χ^2-testin avulla, onko tuotteiden löydettävyydessä eroja kanta-asiakkaiden (osuuskaupan jäsenten) ja muiden asiakkaiden välillä. Ristiintaulukoidaan siis muuttujat 5 ja 7.

Asetetaan nollahypoteesi H_0: Kanta-asiakkaat ja muut asiakkaat ovat samaa mieltä tuotteiden löydettävyydestä.

Chi-square Contingency Table Test for Independence

		1	2	3	4	5	Total
1	Observed	**2**	**4**	**12**	**26**	**9**	53
	Expected	8,48	13,78	9,54	15,90	5,30	53,00
2	Observed	**14**	**22**	**6**	**4**	**1**	47
	Expected	7,52	12,22	8,46	14,10	4,70	47,00
Total	Observed	16	26	18	30	10	100
	Expected	16,00	26,00	18,00	30,00	10,00	100,00

45,80	chi-square
4	df
2,71E-09	p-value

Taulukosta käy selvästi ilmi, että muut kuin kanta-asiakkaat pitivät tuotteiden löydettävyyttä heikkona (rivimuuttujan 2 arvot). Heistä 36 (76,6 %) oli kysymyksessä 7 esitetystä väitteestä joko täysin tai jokseenkin eri mieltä (sarakemuuttujan arvot 1 ja 2). Vaikka yksi odotetuista frekvensseistä on hieman pienempi kuin 5 (4,70), χ^2-testin edellytykset voidaan katsoa voimassa oleviksi (katso kohtaa 7.4).

Testimuuttujan arvo on 45,80 ja vapausasteet 4. Nollahypoteesi hylätään kaikilla merkitsevyystasoilla, koska p-arvo on $2{,}71 \cdot 10^{-9}$ eli käytännössä 0.

(χ^2-jakauman kriittinen arvo 0,1 %:n merkitsevyystasolla vapausastein 4 on 18,467).

Muilla kuin kanta-asiakkailla on vaikeuksia tuotteiden löytämisessä. Tulos on tilastollisesti **erittäin merkitsevä**.

Hintojen esillä oleminen

Kuluttajansuojalain mukaan tuotteiden hintojen tulee olla esillä siten, että asiakkaat voivat vertailla hintoja. Tässäkin tapauksessa voidaan olettaa, että vakioasiakkaat löytävät hintamerkinnät helpommin kuin muut asiakkaat. Tutkitaan asiaa samalla tavalla kuin edellä.

Asetetaan hypoteesi H_0: Kaikki asiakkaat ovat samaa mieltä hintojen esillä olosta, eli kanta-asiakkaat ja muut asiakkaat löytävät hintamerkinnät yhtä hyvin.

Chi-square Contingency Table Test for Independence

		1	2	3	4	5	Total
1	Observed	**3**	**14**	**13**	**13**	**10**	53
	Expected	7,42	19,08	10,60	9,01	6,89	53,00
2	Observed	**11**	**22**	**7**	**4**	**3**	47
	Expected	6,58	16,92	9,40	7,99	6,11	47,00
Total	Observed	14	36	20	17	13	100
	Expected	14,00	36,00	20,00	17,00	13,00	100,00

16,38	chi-square
4	df
,0025	p-value

Huomataan, että muilla asiakkailla on enemmän vaikeuksia hintamerkintöjen löytämisessä kuin kanta-asiakkailla.

Testimuuttujan arvo on 16,38, vapausasteet 4 ja p-arvo 0,0025 eli erehtymisriski on 0,25 %, jos nollahypoteesi hylätään. Se siis hylätään 5 %:n ja 1 %:n merkitsevyystasoilla, mutta sitä ei voida hylätä 0,1 %:n merkitsevyystasolla (kriittinen arvo on 18,467).

Muilla asiakkailla on enemmän vaikeuksia hintamerkintöjen löytämisessä kuin kanta-asiakkailla. Tulos on tilastollisesti **merkitsevä**.

Liikkeen aukioloajat

Kyselyn vastauksissa ilmeni jonkin verran tyytymättömyyttä liikkeen aukioloaikoihin. Tutkitaan, onko tyytyväisyydessä eroja eri ikäryhmien ja eri sukupuolten välillä. Muuttuja 2 (vastaajan ikä) saa kuusi erilaista arvoa ja muuttuja 15 (tyytyväisyys aukioloaikoihin) viisi erilaista arvoa. Tällöin ristiintaulukoitaessa useat solujen teoreettisista frekvensseistä ovat lukua 5 pienempiä eivätkä χ^2-testin edellytykset ole voimassa, mikä voidaan välttää yhdistelemällä luokkia.

Tehdään muuttujille 2 ja 15 uusi luokitus.

Muuttujan 2 uusi luokitus olkoon seuraava:
1 ja 2 = alle 30 vuoden ikäiset
3 ja 4 = 30–49-vuotiaat
5 ja 6 = vähintään 50-vuotiaat.

Muuttuja 15 luokitellaan seuraavasti:
1 ja 2 = eri mieltä olevat
3 = neutraalit
4 ja 5 = samaa mieltä olevat.

Tällöin saadaan seuraava taulukko:

Chi-square Contingency Table Test for Independence

		eri mieltä	neutraalit	samaa mieltä	Total
–29	Observed	**18**	**11**	**3**	32
	Expected	15,04	9,92	7,04	32,00
30–49	Observed	**20**	**13**	**1**	34
	Expected	15,98	10,54	7,48	34,00
50–	Observed	**9**	**7**	**18**	34
	Expected	15,98	10,54	7,48	34,00
Total	Observed	47	31	22	100
	Expected	47,00	31,00	22,00	100,00

29,25	chi-square
4	Df
6,95E-06	p-value

Taulukosta on selvästi havaittavissa, että tyytyväisiä aukioloaikoihin ovat lähinnä vanhimman ikäryhmän edustajat. Kahden nuorimman ikäryhmän edustajista 36 (54,5 %) on tyytymättömiä aukioloaikoihin. Tyytyväisiä näissä ikäryhmissä on vain 6,1 %. Sen sijaan vanhimman ikäryhmän edustajista 52,9 % on tyytyväisiä aukioloaikoihin. Tulos on varmistettavissa χ^2-testillä.

Nollahypoteesi (H_0: Eri-ikäiset ihmiset ovat yhtä tyytyväisiä liikkeen aukioloaikoihin) hylätään kaikilla merkitsevyystasoilla (p-arvo on $6{,}95 \cdot 10^{-6}$).

Nuoret ja keski-ikäiset ovat selvästi tyytymättömämpiä liikkeen aukioloaikoihin kuin yli 50-vuotiaat. Tulos on tilastollisesti **erittäin merkitsevä**.

Tutkitaan vielä, ovatko naiset ja miehet yhtä tyytyväisiä aukioloaikoihin. Pidetään muuttujan 15 luokitus samana kuin edellä. Asetetaan nollahypoteesi H_0: Tyytyväisyys liikkeen aukioloaikoihin on riippumaton sukupuolesta.

Chi-square Contingency Table Test for Independence

		eri mieltä	neutraalit	samaa mieltä	Total
Naiset	Observed	**28**	**24**	**14**	66
	Expected	31,02	20,46	14,52	66,00
Miehet	Observed	**19**	**7**	**8**	34
	Expected	15,98	10,54	7,48	34,00
Total	Observed	47	31	22	100
	Expected	47,00	31,00	22,00	100,00

2,72	chi-square
2	df
,2565	p-value

Tyytyväisyys liikkeen aukioloaikoihin ei näytä riippuvan sukupuolesta. Testimuuttujan arvo on 2,72, vapausasteet 2 ja p-arvo 0,2565. Jos nollahypoteesi hylättäisiin, erehtymisriski olisi yli 25 % (χ^2-jakauman kriittinen arvo 5 %:n merkitsevyystasolla vapausastein 2 on 5,991). Ei siis voida osoittaa, että jompikumpi sukupuoli olisi toista tyytyväisempi liikkeen aukioloaikoihin.

Tiivistelmä tutkimuksesta toimenpidesuosituksineen

Tutkimuksen otos, joka käsitti 100 tilastoyksikköä, on pienehkö varmojen johtopäätösten tekemiseksi. Palvelutasoa mittaavien muuttujien (6–15) jakaumat kuitenkin osoittavat selvästi, mitkä palvelutekijät ovat kunnossa ja mihin kaivataan parannusta. Tutkimuksen päätarkoitus on siten saavutettu. Pieni otos vaikeuttaa johtopäätösten tekemistä asiakaskunnan rakenteesta. Muuttujien 1–5 jakaumista se voidaan kuitenkin likimäärin päätellä.

Tavaratalon asiakaskunta on **naisvaltaista**, ja asiakkaiden kotitaloudet ovat keskimäärin **melko hyvätuloisia**. Asiakaskunta koostuu **kaikenikäisistä henkilöistä**, joista noin puolet on **osuuskaupan jäseniä**. **Lapsiperheitä** asiakaskunnasta on noin puolet.

Tavaratalon vahvuudet palvelutasossa ovat
- myymälän viihtyisyys
- henkilökunnan ystävällisyys ja asiantuntemus
- liikkeessä asioinnin nopeus
- hyvä tuotevalikoima.

Vahvuudet osoittavat, että tavaratalon henkilökunnan koulutus on ajan tasalla ja viihtyisyyteen ja tuotevalikoimaan on palvelumyymäläksi profiloitumisen mukaisesti kiinnitetty runsaasti huomiota. Profiilin mukaisesti tavaratalo ei myöskään kilpaile edullisilla hinnoilla tai usein vaihtuvilla erikoistarjouksilla. Palvelutason heikkoudet taas liittyvät ulkoisiin seikkoihin, jotka lienevät helposti korjattavissa.

Parantamisen varaa on
- liikkeen aukioloajoissa
- tuotteiden löydettävyydessä
- hintojen esillä olemisessa.

Liikkeen aukioloaikoja pitäisi muuttaa siten, että liike palvelisi entistä paremmin niin sanottua aktiiviväestöä. **Aukioloaikoja tulisi muuttaa välittömästi, koska yli puolet alle 50-vuotiaista oli tyytymättömiä niihin.**

Tuotteiden heikko löydettävyys ja hintojen huono esilläolo kiusasivat sellaisia, jotka eivät ole kanta-asiakkaita. **Koska yli 70 % muista asiakkaista oli tätä mieltä, nämä heikkoudet pitää korjata.** Satunnaisesta asiakkaasta, joka ei ilman etsimistä löydä haluamaansa tuotetta ja joka ei voi vertailla kilpailevien tuotteiden hintoja, ei helposti tule kanta-asiakasta.

Jos tavaratalo toteuttaa edellä mainitut toimet, sen tavoite olla parhaiten asiakkaita palveleva tavaratalo kaupungissa on ainakin lähellä toteutumista.

TEHTÄVIÄ

1. Satunnaismuuttuja $X \sim N(\mu, 6)$. Laske odotusarvolle väliestimaatti
a) 95 %:n,
b) 99 %:n,
c) 99,9 %:n
luottamustasolla, kun 50 kappaleen otoksesta laskettu keskiarvo on 25.

2. Korkeakoulussa tutkitaan näyttöpäätetyöskentelyn haittoja ottamalla 100 päivittäin näyttöpäätteen äärellä työskentelevän henkilön otos. Otokseen joutuneiden henkilöiden keskimääräinen poissaoloaika työstä oli vuoden aikana 96 tuntia. Poissaolotuntien keskihajonta oli 32 tuntia. Määritä keskimääräiselle poissaololle
a) 95 %:n
b) 99,9 %:n luottamusväli.

3. Auton moottorin männän halkaisijan keskiarvoksi on saatu 180 kappaleen otoksesta 89,967 mm. Samasta otoksesta laskettu keskihajonta on 0,003 mm. Määritä männän halkaisijalle 99 %:n luottamusväli.

4. Laaduntarkkailussa tutkittiin valmistuneiden paristojen laatua ottamalla 50 pariston otos. Otoksesta laskettiin paristojen keskimääräinen jännite ja jännitteiden keskihajonta, jolloin tulokseksi saatiin $\overline{x} = 4{,}56\ V$ ja $s = 0{,}1\ V$. Laske 95 %:n väliestimaatti paristojen keskimääräiselle jännitteelle.

5. Pesukoneita valmistava yritys halusi tutkia, kuinka pitkä takuuaika heidän valmistamilleen tuotteille voitaisiin myöntää. Valmistuneista pesukoneista otettiin 50 kappaleen otos. Laitteita testattaessa todettiin, että niihin tuli ensimmäinen vika keskimäärin 775 käyttötunnin jälkeen. Vastaava keskihajonta oli 200 tuntia. Kuinka monen kuukauden takuu pesukoneille voidaan myöntää, jotta takuukorjaukset voitaisiin välttää 95 %:n varmuudella, kun pesukoneiden päivittäiseksi käyttöajaksi arvioidaan 2 tuntia?

6. Yritys tutki valmistamiensa tekstiilien pesunkestävyyttä. Tutkimusta varten tuotantoerästä poimittiin 150 kappaleen satunnaisotos. Poimittujen tuotteiden pesunkestävyys oli keskimäärin 23,5 pesukertaa ja keskihajonta 4,2 pesukertaa. Myöhemmin tekstiilien raaka-aineen koostumusta muutettiin. Kun sen jälkeen testattiin pesunkestävyyttä, 150 otokseen joutuneen tekstiilin pesunkestävyys oli keskimäärin 26,2 pesukertaa ja keskihajonta 4,0 pesukertaa. Tutki väliestimaattien avulla, voidaanko 95 %:n todennäköisyydellä päätellä, että tekstiilien pesunkestävyys parantui, kun niiden raaka-aineen koostumusta muutettiin.

7. Tuote-erästä otetussa 2 000 kappaleen otoksessa löydettiin 126 viallista tuotetta. Määritä viallisten tuotteiden suhteelliselle osuudelle väliestimaatti 99 %:n luottamustasolla.

8. Edellisen tehtävän koko valmistuserä oli 66 800 kappaletta. Määritä viallisten tuotteiden määrälle väliestimaatti 95 %:n luottamustasolla.

9. Puhto-pesujauheen markkinaosuutta tutkittaessa havaittiin, että 570 satunnaisesti valitusta taloudesta 115 käytti Puhtoa. Laske 95 %:n luottamusväli pesujauheen markkinaosuudelle.

10. Vakuutusyhtiö myöntää vapaa-ajan tapaturmavakuutuksia. Vakuutuksen perusteella maksetaan 1 200 euron kertakorvaus tapaturmasta, joka aiheuttaa vähintään kuukauden kestävän työkyvyttömyyden. Vuonna 20XX vakuutuksen oli ottanut 73 200 henkilöä ja korvaus maksettiin 2 100 henkilölle.

a) Laske 99 %:n luottamusväli seuraavana vuonna maksettavien korvausten lukumäärälle, kun vakuutuksenottajien määrän arvioidaan kasvavan 8 %.

b) Mikä on vakuutusmaksun suuruus seuraavana vuonna, kun vakuutusmaksutulon on ylitettävä 30 prosentilla maksettujen korvausten enimmäismäärä?

11. Mikrotietokoneiden maahantuoja toi eräänä vuonna maahan 3 500 mikrotietokonetta. Näistä 150:lle jouduttiin tekemään takuukorjauksia. Seuraavana vuonna samoja mikrotietokoneita aiotaan tuoda 70 % edellisvuotta enemmän. Määritä 95 %:n luottamusväli jälkimmäisenä vuonna takuun perusteella korjattavien mikrotietokoneiden määrälle.

12. Tavaratalo on arvioinut, että hävikki (varkaudet, pilaantuminen ym.) on noin 1,2 % liikevaihdosta. Arvio perustuu tilastoon, jossa on 98 tavaratalon hävikit 10 vuoden ajalta. Talousjohtajan tekemän budjetin mukaan vuoden 20XX arvioitu liikevaihto on 65 miljoonaa euroa.

a) Määritä vuoden 20XX hävikille (milj. €) 95 %:n luottamusväli.

b) Hävikki otetaan huomioon hinnoittelussa siten, että hintoihin lisätään hävikkiprosentin enimmäismäärä 95 %:n luottamustasolla. Jos asiakas ostaa tavaratalosta tuotteita 500 eurolla kuukaudessa, kuinka monta euroa ylimääräistä hän joutuu hävikin vuoksi maksamaan vuosittain?

13. Kilpailevien elintarvikkeiden markkinaosuutta tutkittiin valitsemalla satunnaisesti 600 suomalaisen henkilön otos. Näistä henkilöistä 215 ilmoitti käyttävänsä tuotetta A ja 198 tuotetta B. Määritä molempien tuotteiden markkinaosuuksille 95 %:n luottamusvälit ja päättele niiden avulla, voidaanko tuotteen A markkinaosuutta Suomessa pitää suurempana kuin tuotteen B markkinaosuutta.

14. Kuinka suuri otoskoon pitäisi vähintään olla, jotta edellisessä esimerkissä tuotteen A markkinaosuutta Suomessa voitaisiin 95 %:n varmuudella pitää suurempana kuin tuotteen B markkinaosuutta, kun otoksesta lasketut suhteelliset osuudet ovat seuraavat 35,8 % (A) ja 33,0 % (B).

15. Tietokoneissa käytettävien levyjen valmistaja ilmoittaa levyn uran kestävän keskimäärin 30 miljoonaa lukukertaa. Kun testattiin 500 levyä, joista kullakin oli 80 uraa, havaittiin, että ura kestää keskimäärin 29 960 000 lukukertaa. Keskihajonta oli 5 miljoonaa lukukertaa. Testaa 5 %:n merkitsevyystasolla, onko levyjen laatu huonontunut.

16. Tulitikkurasiassa on tehtaan ilmoituksen mukaan keskimäärin 45 tulitikkua. Timo Tilastoija ostaa vuoden aikana 100 rasiaa kyseisen tehtaan tulitikkuja ja laskee jokaisessa rasiassa olevan tulitikkujen määrän. Hän saa tulitikkujen määräksi keskimäärin 44 tikkua rasiaa kohden ja rasioissa olevien tikkujen määrän keskihajonnaksi kaksi tulitikkua. Vuoden lopussa Timo kirjoittaa tehtaalle ja kertoo, ettei tulitikkurasioissa ole luvattua määrää tulitikkuja. Tehtaan tiedottaja ilmoittaa Timolle, että tämä on erehtynyt, koska tämän otos ei ollut riittävän suuri johtopäätöksen tekemiseksi. Testaa 0,1 %:n merkitsevyystasolla, onko Timon väite oikea.

17. Leipomo ilmoittaa, että sen leipomat leivät painavat keskimäärin 500 g. Kuluttajatutkimuksessa 50 leivän otoksesta laskettu leipien keskimääräinen paino oli 480 g ja painojen hajonta 65 g. Testaa 1 %:n merkitsevyystasolla hypoteesi H_0: $\mu = 500$ g käyttämällä
a) kaksisuuntaista,
b) yksisuuntaista testiä.

18. Poimitaan 10 tilastoyksikön riippumaton otos normaalisti jakautuneesta perusjoukosta. Otoksesta laskettu keskiarvo on 39,7 ja keskihajonta 6,1. Asetetaan hypoteesit H_0: $\mu = 35$; H_1: $\mu \neq 35$. Testaa 5 %:n ja 1 %:n merkitsevyystasoilla, onko nollahypoteesi hylättävä.

19. Yritys ilmoitti valmistamansa kasvisrasvan markkinaosuudeksi 13,5 %. Kilpailija tutki väitteen poimimalla eri puolilta maata 12 myymälän otoksen. Niissä kyseisen rasvan keskimääräinen markkinaosuus oli 12,2 % ja markkinaosuuksien hajonta 3,6 %. Testaa 5 %:n merkitsevyystasolla yksisuuntaista testiä käyttämällä, voidaanko valmistajan ilmoitus kumota.

20. Veneen ankkuriköyden vetolujuudeksi ilmoitettiin 15 000 N. Kun väitteen paikkansapitävyyttä testattiin, tehtiin viisi kertaa koe, jossa köyttä kuormitettiin, kunnes se katkesi. Keskimääräiseksi vetolujuudeksi näissä kokeissa saatiin 14 520 N ja vetolujuuksien hajonnaksi 1 650 N. Testaa 5 %:n merkitsevyystasolla yksisuuntaista testiä käyttämällä hypoteesi H_0: $\mu = 15\,000$ N.

21. Autotehdas ilmoittaa, että se on parantanut valmistamiensa autojen jarruja, minkä ansiosta auton jarrutusmatka 100 kilometrin tuntinopeudesta on lyhentynyt 50 metristä 8 %. Tehtaan koeajaja testasi jarruja kuudella satunnaisesti valitulla autolla. Tällöin keskimääräiseksi jarrutusmatkaksi saatiin 48,3 metriä ja jarrutusmatkojen keskihajonnaksi 2,0 metriä. Testaa

a) 5 %:n,

b) 1 %:n merkitsevyystasolla, onko valmistajan ilmoitus oikea.

22. Asetetaan tehtävässä 6 nollahypoteesi H_0: Tekstiilien pesunkestävyys ei ole parantunut. Testaa yksisuuntaista testiä käyttämällä hypoteesi 1 %:n merkitsevyystasolla.

23. Tutkimuksessa selvitettiin, poikkeaako kaupungeissa asuvien ihmisten alkoholinkäyttö maalla asuvien alkoholinkäytöstä. Sekä kaupungeissa että maalla asuvista poimittiin 50 henkilön satunnaisotos. Kaupungeissa asuvat käyttivät kuukauden aikana alkoholia keskimäärin 0,65 litraa (muutettuna 100-prosenttiseksi alkoholiksi) ja kulutuksen keskihajonta oli 0,25 litraa. Vastaavat tunnusluvut maalla asuville olivat 0,50 litraa ja 0,22 litraa. Testaa 1 %:n merkitsevyystasolla hypoteesi H_0: Kaupungeissa ja maalla asuvat käyttävät keskimäärin yhtä paljon alkoholia.

24. Tutkimuksessa pyrittiin selvittämään, käyttävätkö työttömät enemmän kirjastojen palveluja kuin työssäkäyvät. Työttömistä poimittiin 120 hengen ja työssä käyvistä 85 hengen otos. Otokseen joutuneet työttömät kävivät kuukauden aikana kirjastossa keskimäärin 5,3 kertaa, ja keskihajonta oli 2,5 kertaa. Työssä kävijöiden otoksesta lasketut vastaavat tunnusluvut olivat 3,7 kertaa ja 2,4 kertaa. Testaa 0,1 %:n riskitasolla hypoteesi H_0: Työttömät ja työssä käyvät käyttävät kirjastojen palveluja yhtä paljon.

25. Kun tutkittiin taksiautoilijoiden ja myyntimiesten vuotuista ajokilometrien määrää, poimittiin satunnaisesti 43 taksiautoilijan ja 40 myyntimiehen otokset. Heidän vuotuiset ajokilometrimääränsä vaihtelivat seuraavasti:

Taksiautoilijat (1 000 km): 67; 58; 86; 50; 75; 82; 49; 66; 42; 58; 77; 44; 56; 59; 88; 40; 59; 61; 60; 44; 92; 55; 50; 43; 36; 77; 54; 49; 66; 68; 61; 71; 43; 48; 59; 61; 40; 48; 68; 61; 73; 51; 47.

Myyntimiehet (1 000 km): 68; 53; 81; 72; 69; 60; 52; 59; 51; 45; 87; 48; 59; 62; 78; 66, 59; 50; 51; 91; 77; 71; 43; 39; 47; 34; 50; 48; 89; 73; 76; 80; 67; 52; 51; 73; 65; 60; 51; 70.

Testaa 5 %:n merkitsevyystasolla hypoteesi H_0: Taksiautoilijoiden ja myyntimiesten vuotuisten ajokilometrien määrät eivät poikkea toisistaan.

26. Kunta tutki asukkaidensa terveyspalvelujen käyttöä. Haluttiin selvittää muun muassa se, käyttävätkö eläkeläiset enemmän terveyspalveluja kuin työikäinen väestö. Asian selvittämiseksi poimittiin 20 työikäisen ja 16 eläkeläisen satunnaisotokset. Heidän käyntimääränsä terveyskeskuksessa vuoden 20XX aikana olivat seuraavat:

Työikäiset: 2, 0, 6, 1, 8, 4, 3, 2, 5, 4, 5, 3, 1, 7, 6, 5, 4, 5, 5, 4

Eläkeläiset: 7, 8, 3, 6, 6, 7, 3, 4, 3, 5, 8, 6, 9, 2, 4, 5.

Testaa 5 %:n merkitsevyystasolla yksisuuntaista testiä käyttämällä hypoteesi H_0: Työikäiset ja eläkkeellä olevat kuntalaiset käyttävät keskimäärin yhtä paljon terveyskeskuksen palveluja, kun oletetaan, että terveyskeskuskäyntien määrä molemmissa perusjoukoissa noudattaa normaalijakaumaa.

27. Kun vertailtiin ekonomien ja insinöörien ansiotasoa suuryrityksissä 10 vuotta valmistumisen jälkeen, poimittiin 20 insinöörin ja 20 ekonomin satunnaisotokset. Otoksiin joutuneiden palkat (€/kk) olivat seuraavat:

Ekonomit: 2 800; 4 400; 3 700; 6 000; 3 800; 5 800; 2 700; 7 100; 5 500; 5 000; 6 300; 5 800; 3 500; 3 600; 6 200; 4 600; 4 100; 6 600; 5 300; 4 700.

Insinöörit: 3 600: 4 900; 4 700; 2 900; 5 400; 5 500; 6 300, 6 700; 4 800; 3 900; 6 000; 7 300; 5 300; 5 100; 4 900; 3 900; 5 400; 3 700; 4 600; 4 900.

Testaa 5 %:n merkitsevyystasolla kaksisuuntaista testiä käyttäen hypoteesi H_0: Ekonomien ja insinöörien keskimääräiset palkat eivät poikkea toisistaan, kun oletetaan, että heidän palkkansa perusjoukoissa noudattavat normaalijakaumaa.

28. Tutkimuksessa testattiin, oliko uudella lisäravinteella vaikutusta kestävyysurheilijoiden veren hemoglobiiniarvoihin. Tutkimusta varten poimittiin 15 kestävyysurheilijan satunnaisotos. Otokseen joutuneiden veren hemoglobiiniarvot mitattiin, minkä jälkeen heille annettiin kahden kuukauden ajan kyseistä lisäravinnetta. Testattujen urheilijoiden veren hemoglobiiniarvot olivat seuraavat.

Ennen: 136, 129, 140, 137, 128, 145, 156, 141, 133, 126, 122, 147, 138, 130, 154.

Jälkeen: 142, 128, 145, 138, 134, 152, 165, 145, 141, 134, 128, 149, 136, 135, 154.

Testaa 1 %:n merkitsevyystasolla yksisuuntaisella verrannollisten parien t-testillä hypoteesi H_0: Lisäravinne ei vaikuta kestävyysurheilijoiden veren hemoglobiinipitoisuuteen.

29. Laaduntarkkailussa havaittiin aikaisemmin, että viallisten tuotteiden suhteellinen osuus tuotannosta oli keskimäärin 5,2 %. Yritys päättää parantaa tuotantomenetelmiään viallisten osuuden vähentämiseksi. Uutta tuotantomenetelmää testattaessa havaitaan, että 1 850 tuotteesta 73 on viallisia. Testaa

a) 5 %:n,

b) 1 %:n merkitsevyystasolla hypoteesi H_0: Tuotteiden laatu ei ole parantunut.

30. Alueen 50 vuotta täyttäneistä asukkaista 42 % kärsi liian korkeasta verenpaineesta. Asukkaille päätettiin ryhtyä antamaan terveysvalistusta. Kun 10 vuoden kuluttua tutkittiin terveysvalistuksen vaikutusta, havaittiin, että alueelta satunnaisesti valitusta 260 yli 50-vuotiaasta 72 kärsi liian korkeasta verenpaineesta. Testaa

a) 1 %:n,

b) 0,1 %:n merkitsevyystasolla hypoteesi H_0: Terveysvalistuksesta ei ole apua korkeasta verenpaineesta kärsiville.

31. Vuonna 20XX 23,7 % Suomen kotitalouksista oli erään lehden tilaajia. Lehtiyhtiö päätti tehostaa lehden markkinointia tilaajamäärän kasvattamiseksi, minkä ansiosta lehden tilaajien määrä kasvoi seuraavana vuonna 0,5 prosenttiyksikköä. Levikkitutkimuksessa Itä-Suomessa haastatelluista 242 kotitaloudesta 55 oli tuolloin lehden tilaajia. Testaa 5 %:n merkitsevyystasolla hypoteesi H_0: Lehden tilaajien suhteellinen osuus Itä-Suomessa ei poikkea koko Suomessa lehden tilanneiden suhteellisesta osuudesta.

32. On havaittu, että lääke A lieventää oireita 63 potilaalla sadasta. Uutta lääkettä B testattaessa puolestaan havaittiin, että 1 230 tutkitusta potilaasta 805 oli saanut lääkkeestä lievitystä oireisiinsa. Testaa 1 %:n riskitasolla hypoteesi H_0: Uuden ja vanhan lääkkeen teho on sama. Käytä yksisuuntaista testiä.

33. Tutkimuksen mukaan keväällä 20XX ydinvoiman lisärakentamista vastusti 42,1 % suomalaisista. Kantaansa ei osannut sanoa 19,5 %. Myöhemmin tehdyssä kyselyssä 1 440 satunnaisesti valitusta haastatellusta 570 kannatti ydinvoiman lisärakentamista. Kantaansa ei osannut sanoa 275 henkilöä. Testaa 5 %:n merkitsevyystasolla hypoteesi H_0: Ydinvoiman lisärakentamista vastustavien osuus ei ole muuttunut.

34. Vakuutusyhtiö Väinämöinen tutki vapaaehtoisen tapaturmavakuutuksensa markkinaosuutta eteläisellä ja pohjoisella alueella. Eteläisellä alueella haastateltiin 860 vakuutuksen ottajaa, joista 25,1 % oli ottanut vakuutuksensa Väinämöisestä. Pohjoisella alueella haastatellusta 475 vakuutuksen ottajasta 128 oli Väinämöisen asiakkaita. Testaa 1 %:n merkitsevyystasolla hypoteesi H_0: Vakuutusyhtiön markkinaosuus molemmilla alueilla on sama.

35. Työttömyyden aiheuttamien ongelmien tutkimiseksi haastateltiin 1716 satunnaisesti valittua työtöntä. Haastatelluista 48,6 % oli naisia. Miehistä 573 ja naisista 500 ilmoitti kärsivänsä masennuksesta. Testaa 0,1 %:n merkitsevyystasolla yksisuuntaista testiä käyttämällä hypoteesi H_0: Masennuksesta kärsivien suhteellinen osuus molemmissa ryhmissä on sama.

36. Testaa 5 %:n merkitsevyystasolla tehtävän 26 hypoteesi H_0: Työikäiset ja eläkkeellä olevat kuntalaiset käyttävät keskimäärin yhtä paljon terveyskeskuksen palveluja, kun tehtävän normaalijakaumaolettamus ei ole voimassa.

37. Testaa 5 %:n merkitsevyystasolla tehtävän 27 hypoteesi, kun normaalijakaumaolettamus ei ole voimassa.

38. Testaa 5 %:n merkitsevyystasolla tehtävän 28 hypoteesi, kun verrannollisten parien *t*-testin oletukset (normaalijakaumaolettamus) eivät ole voimassa.

39. Kun vertailtiin erään kaupungin työntekijöiden ja kaupungissa toimivien yritysten työntekijöiden ikäjakaumaa, poimittiin molemmista perusjoukoista 12 hengen otokset. Otokseen poimittujen iät olivat seuraavat.

Kaupunki: 38, 47, 44, 56, 26, 59, 47, 45, 37, 61, 42, 45.

Yritykset: 44, 30, 55, 60, 51, 25, 33, 50, 40, 23, 42, 37.

Testaa 5 %:n merkitsevyystasolla hypoteesi H_0: Kaupungin työntekijöiden ja yritysten työntekijöiden ikäjakaumat eivät poikkea toisistaan. Ei voida olettaa, että työntekijöiden iät perusjoukossa noudattavat normaalijakaumaa.

40. Tutkittiin alkoholin vaikutusta henkilöiden tarkkaavaisuuteen. Asian selvittämiseksi poimittiin satunnaisesti 15 koehenkilön ryhmä. Koehenkilöille järjestettiin tarkkaavaisuutta mittaavat kokeet. Kokeissa tehdyt virheet tilastoitiin ennen alkoholin nauttimista ja puoli tuntia sen jälkeen, kun koehenkilöt olivat nauttineet kaksi pullollista keskiolutta. Virheiden määrät kokeissa olivat seuraavat.

Ennen: 2, 5, 3, 7, 6, 4, 11, 8, 10, 1, 3, 5, 8, 4, 7.

Jälkeen: 5, 5, 4, 10, 7, 8, 9, 9, 12, 5, 2, 6, 11, 4, 9.

Testaa 1 %:n merkitsevyystasolla hypoteesi H_0: Kahden keskioluen nauttiminen ei vaikuta ihmisten tarkkaavaisuuteen. Havaintoarvojen erotusten ei voida olettaa noudattavan normaalijakaumaa.

41. Tilastotieteilijä teki tietokoneohjelman, joka arpoo satunnaislukuja. Ohjelmaa testatessaan hän halusi selvittää, esiintyvätkö kaikki numerot yhtä monta kertaa luvuissa, jos lukuja arvotaan riittävän paljon. Kun hän arpoi 100 kolmenumeroista lukua, numeroiden jakauma oli seuraava:

Numero	0	1	2	3	4	5	6	7	8	9
Lukumäärä	23	26	33	24	35	28	31	36	27	37

Arvioi χ^2-testin avulla, onko ohjelma luotettava eli voidaanko päätellä, että kaikki numerot esiintyvät yhtä monta kertaa, kun lukuja arvotaan riittävän paljon. Merkitsevyystaso on 1 %.

42. Kansanedustaja Tomeran saamista äänistä tuli 45 % naisilta ja loput miehiltä. Ennen seuraavia vaaleja tehdyssä tutkimuksessa poimittiin vaalipiirin äänioikeutetuista 1 200 henkilön otos. Otokseen valituista 24,5 % sanoi äänestävänsä Tomeraa seuraavissa vaaleissa. Heistä 112 oli naisia. Testaa 1 %:n riskitasolla hypoteesi H_0: Tomeran saamista äänistä 45 % tulee naisilta myös uusissa vaaleissa.

43. Markkinoilla olevien keskenään kilpailevien tuotteiden markkinaosuudet olivat vuonna 20XX seuraavat:

Tuote	A	B	C	D
Markkinaosuus	28,5 %	16,7 %	30,3 %	24,5 %

Seuraavana vuonna markkinoille ei tullut uusia kilpailevia tuotteita. Markkinatutkimuksessa havaittiin, että sadasta tämän tuoteryhmän tuotteen ostaneista 24 osti tuotteen A, 18 tuotteen B, 34 tuotteen C ja 24 tuotteen D. Testaa 5 %:n merkitsevyystasolla hypoteesi H_0: Markkinaosuudet eivät ole muuttuneet.

44. Hampurilaisravintola valmistaa neljää erilaista hampurilaista: Juustoa, Ruista, Isoa ja Jättiä. Kysyttäessä asiakkailta parasta hampurilaista saatiin seuraavat tulokset: Juusto 18 ääntä, Ruis 23 ääntä, Iso 35 ääntä ja Jätti 24 ääntä. Voidaanko tuloksista tilastollisin perustein päätellä, että kaikki hampurilaiset eivät ole yhtä suosittuja? Sovella χ^2-testiä 5 prosentin merkitsevyystasolla. (Yo, 2003 K, lyhyt)

45. Nuorison kiinnostusta tietotekniikkaan tutkittiin ottamalla 500 koululaisen otos. Vastausten jakauma oli seuraava:

	Kiinnostunut	Ei kiinnostunut	Yhteensä
Tytöt	160	120	280
Pojat	140	80	220

Testaa 1 %:n merkitsevyystasolla hypoteesi H_0: Kiinnostus tietotekniikkaan on riippumaton sukupuolesta.

46. Tuotteen valmistaja tutki TV-mainonnan vaikutusta tuotteen menekkiin. Ennen mainonnan aloittamista tavaratalon 400 asiakkaasta 80 osti tuotteen. Mainoskampanjan päättymisen jälkeen saman liikkeen asiakkaista otettiin uusi 400 henkilön otos. Henkilöistä 140 osti tuotteen. Testaa 0,1 %:n merkitsevyystasolla hypoteesi H_0: Mainonnalla ei ole vaikutusta tuotteen menekkiin.

47. Yliopisto teki tutkimuksen veren kolesteroliarvojen ja sydänsairauksien välisestä yhteydestä. Koehenkilöinä olivat 50 vuotta täyttäneet miehet. Tulokset olivat seuraavat:

Kolesteroliarvo mmol/l	Sairastui	Ei sairastunut	Yhteensä
–4,0	48	267	315
4,1–5,0	36	196	232
5,1–6,0	87	324	411
6,1–7,0	51	112	163
7,1–	82	128	210

Testaa 1 %:n merkitsevyystasolla hypoteesi H_0: Kolesteroliarvolla ei ole vaikutusta yli 50-vuotiaiden miesten sydänsairauksiin.

8
Tilastollinen riippuvuus

Taloustieteilijät pyrkivät löytämään talouden lainalaisuuksia. Yksi tällainen sääntö on, että osakesijoittamisessa tuotto ja riski kulkevat käsi kädessä. Toisen säännön mukaan hajauttaminen osakesijoittamisessa pienentää riskiä. Kolmas sääntö on, että alhainen korkotaso suosii sijoittamista pörssiin.

Paljon tutkitaan myös muita kuin taloudellisia riippuvuuksia, esimerkiksi tupakoinnin ja keuhkosyöpään sairastuvuuden ja kasvihuonekaasujen ja maapallon lämpenemisen välistä yhteyttä. Talouden lakeja tai yleisiä riippuvuuksia löydetään samalla tavalla kuin vanhan kansan viisauksia aikanaan: seuraamalla ilmiöiden kehittymistä tai niiden esiintymistä samanaikaisesti jonkin toisen ilmiön kanssa ja tekemällä sitten johtopäätöksiä havainnoista.

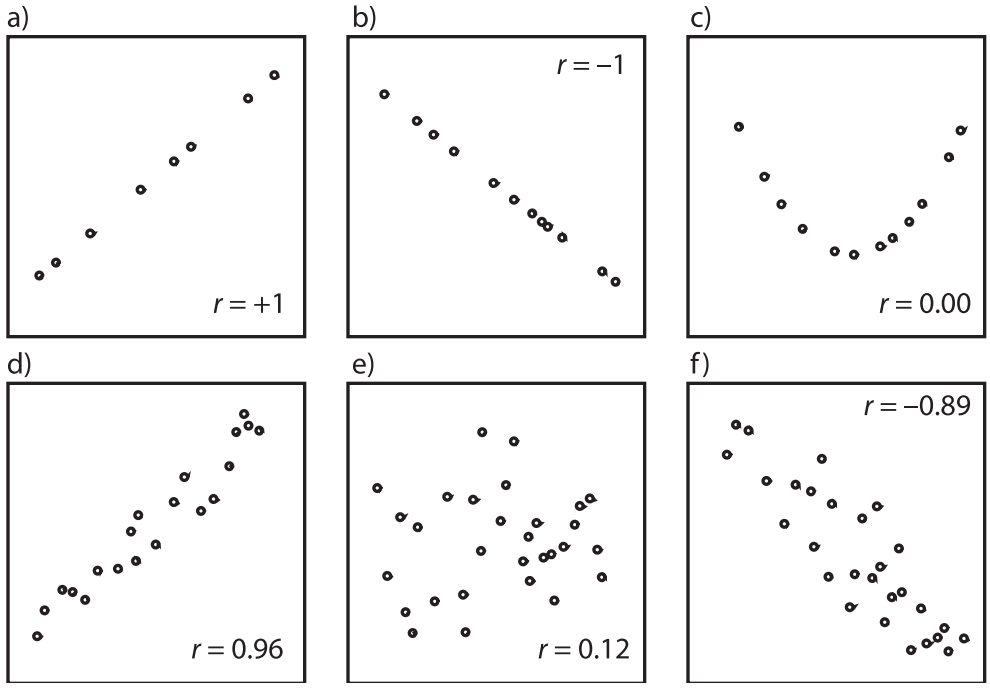
Tässä luvussa tarkastellaan tilastollista riippuvuutta. Käytännössä riippuvuutta tutkitaan siten, että aluksi luodaan aikaisempien havaintojen perusteella työhypoteesi. Sen jälkeen kerätään tarvittava tilastoaineisto ja lopuksi testataan työhypoteesi. Esimerkissä 2 tutkitaan riippuvuutta ja myöhemmin esimerkissä 5 tehdään testi, jossa tutkitaan pitääkö paikkaansa työhypoteesi, jonka mukaan työttömyysaste ja inflaatioaste olisivat kytköksissä toisiinsa.

8.1 Hajontakuviot

Perustapaus

Kahden muuttujan välisen yhteyden tutkiminen aloitetaan **järkevyystarkastelulla**. Aluksi pohditaan, onko tarkasteltavilla muuttujilla mitään mielekästä mahdollisuutta olla yhteydessä toisiinsa. Jos vastaus on kielteinen, tilastoaineiston muu tarkastelu on tässä mielessä turhaa. Toisaalta tarkastelua kannattaa jatkaa, ellei ole mitään loogista tai muuta järkevää estettä sille, että muuttujilla voisi olla jokin yhteys toisiinsa.

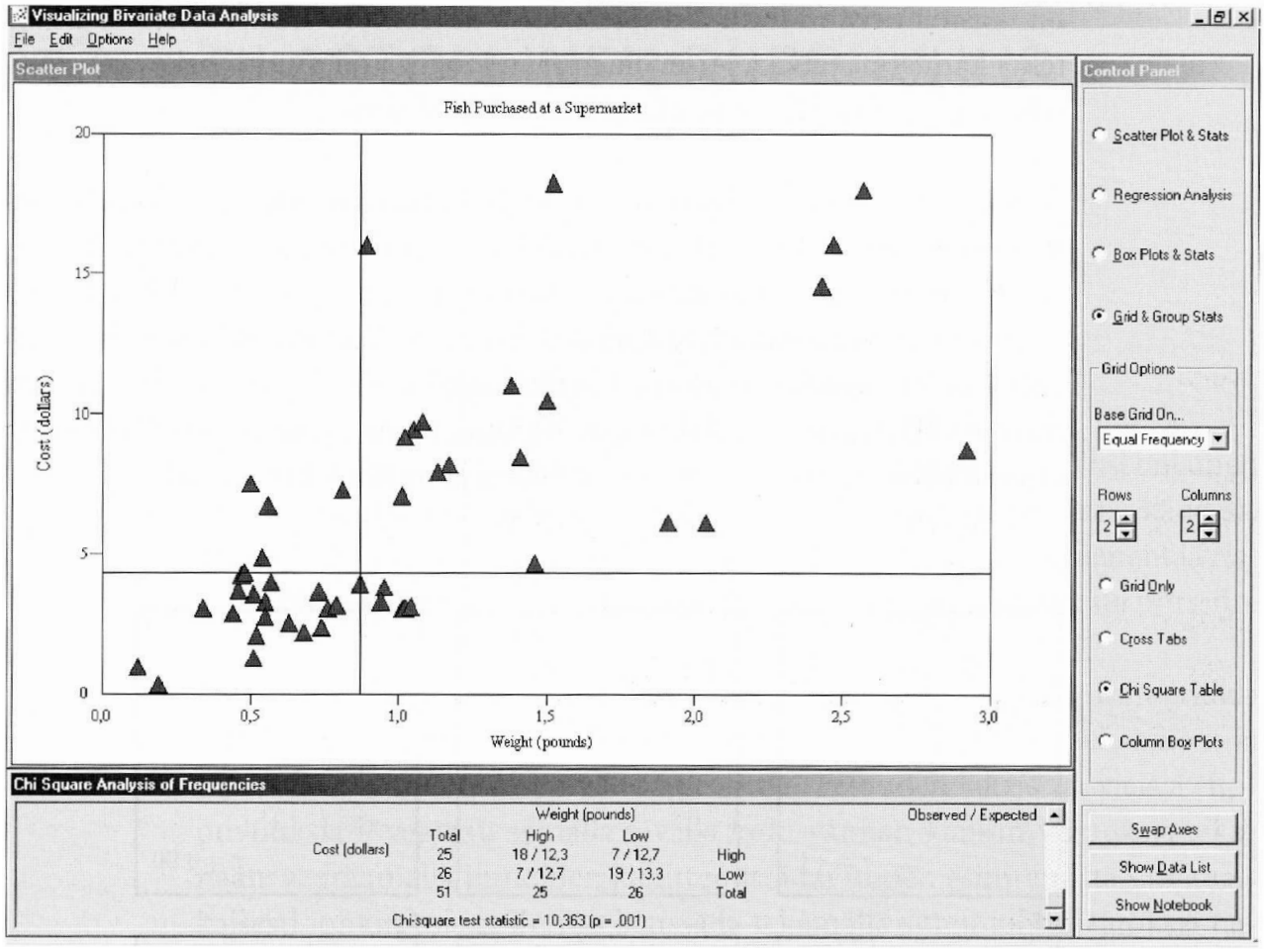
Seuraavaksi tilastoaineistoa tutkitaan **hajontakuvioiden** (scatterplot) eli **korrelaatiodiagrammien** (scatter diagram) avulla. Tätä varten merkitään tarkasteltavien muuttujien arvot koordinaatistoon ja tutkitaan, muodostavatko pisteet jollakin tavalla säännöllisen pistejoukon. Hajontakuviosta käyvät ilmi mahdollisen yhteyden voimakkuus, muoto ja suunta. Käytännössä kuvio tehdään tietokoneohjelmien avulla. Käytetystä ohjelmasta riippuu, kuinka paljon itse voi muokata kuvion ulkomuotoa, asteikkoja, otsikoita ja muita selitystekstejä.



Hajontakuvioita.

Yllä olevan kuvan kolmessa ensimmäisessä kuviossa pisteiden välillä vallitsee funktionaalinen yhteys, sillä kaikki pisteet sijaitsevat joko suoralla tai paraabelilla. Muissa kuvioissa yhteyden voimakkuus ja suunta vaihtelee.

Joissakin tilasto-ohjelmissa on mahdollisuus tarkastella hajontakuviota monipuolisesti, myös interaktiivisesti. Seuraavalla sivulla oleva kuva on Visual Statistics -nimisestä ohjelmasta. Kuvan hajontakuvio on jaettu neljään soluun, joissa ristikkäisissä soluissa on likimäärin sama määrä havaintoja. Solujen lukumäärää ja niiden muodostamisen perustetta voi muutella.



Hajontakuvio Visual Statistics -ohjelmasta.

Samanaikaiset hajontakuviot

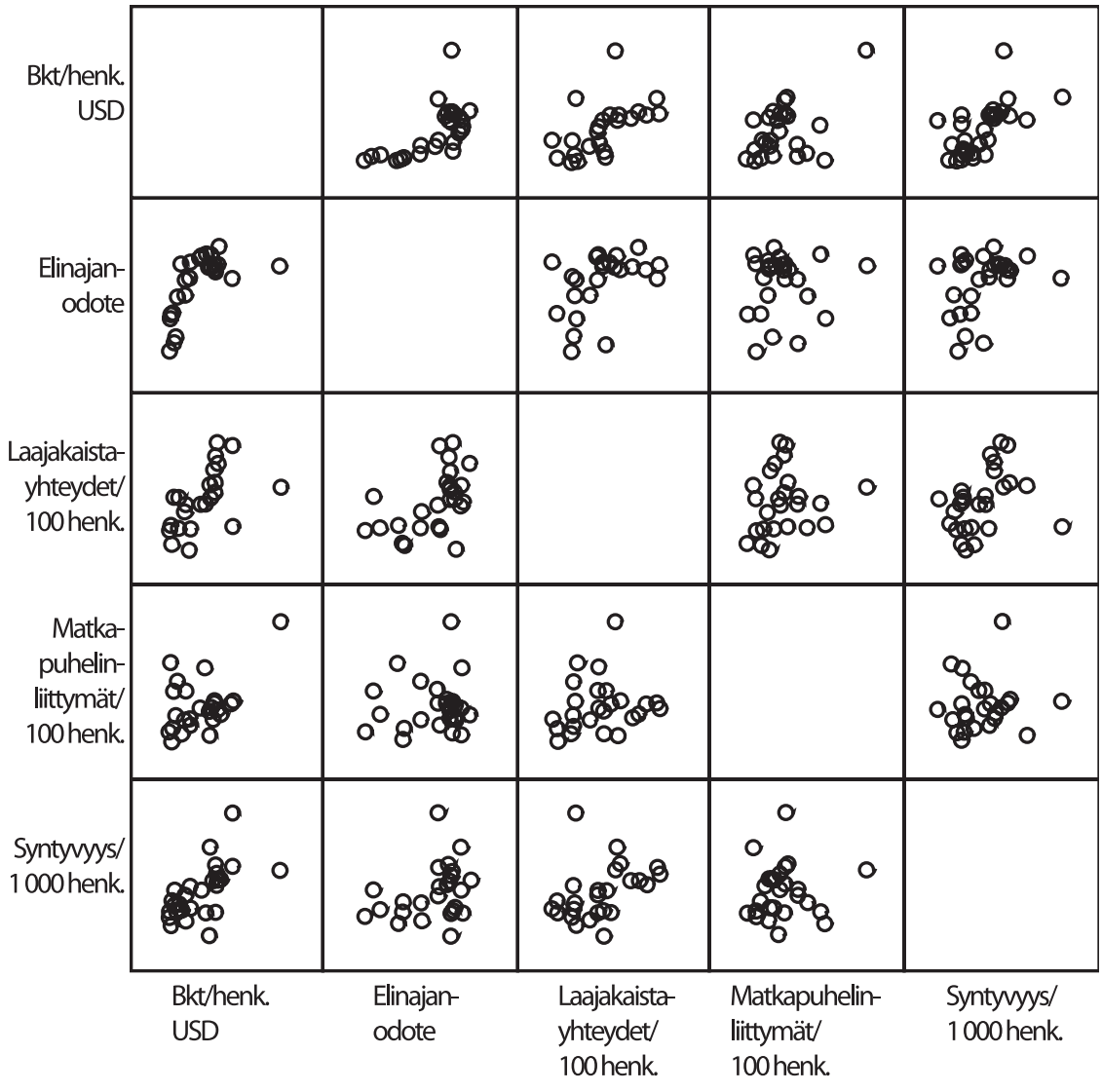
Tarkastelemalla parittaisesti ja samanaikaisesti useaa muuttujaa on mahdollisuus löytää nopeasti lupaavimmat ehdokkaat toisiinsa liittyville muuttujille.

ESIMERKKI 1

Seuraavassa taulukossa on tilastotietoja eräiden EU-maiden osalta vuodelta 2005 (lähde: Tilastokeskus). Taulukossa ”Bkt/henk.” tarkoittaa ostovoimakorjattua bruttokansantuloa henkeä kohti Yhdysvaltojen dollareina. ”Elinajanodote” tarkoittaa odotettavissa olevaa elinaikaa 0-vuotiaalle.

Tilastotietoja eräistä EU-maista

Maa	Bkt/henk. USD	Elinajan-odote	Laajakais-tayhteydet /100 henk.	Matkapuhe-linliittymät /100 henk.	Syntyvyys /1000 henk.
Alankomaat	38 618	78,8	25,2	97,2	11,5
Belgia	35 712	78,6	19,1	90,8	11,2
Britannia	37 023	78,4	16,4	102,2	12,0
Espanja	27 226	79,5	11,5	96,8	10,7
Irlanti	48 604	77,6	6,6	101,5	14,7
Italia	30 200	79,7	11,7	124,3	9,5
Itävalta	37 117	78,9	14,2	99,8	9,5
Kreikka	20 327	79,1	1,4	90,3	9,7
Kypros	20 214	77,7	6,1	86,1	10,9
Latvia	6 862	71,1	5,7	81,1	9,3
Liettua	7 446	74,0	6,9	127,1	8,9
Luxemburg	80 288	78,7	15,3	154,8	11,7
Malta	13 803	78,9	12,7	80,8	9,6
Portugali	17 456	77,5	11,6	109,1	10,4
Puola	7 946	74,4	2,7	75,7	9,5
Ranska	33 918	79,6	15,8	79,4	12,9
Ruotsi	39 694	80,4	20,6	93,3	11,2
Saksa	33 854	78,7	12,8	95,8	8,3
Slovakia	8 775	74,5	2,6	84,1	10,1
Slovenia	16 986	76,1	9,8	89,4	9,1
Suomi	37 504	78,4	22,3	99,7	11,0
Tanska	47 984	77,6	24,6	100,7	11,9
Tsekki	12 152	76,0	6,3	115,2	10,0
Unkari	10 814	72,4	6,2	92,3	9,7
Viro	9 727	71,8	13,3	108,8	10,7



Samanaikaiset hajontakuviot.

Kuvasta käy ilmi, että osa kuvioista esittää melko säännönmukaista pistejoukkoa, kun taas osassa kuvioita mitään selkeää säännönmukaisuutta ei voida havaita.

8.2 Korrelaatio

Ellei hajontakuviossa ilmene minkäänlaista säännönmukaisuutta, on lähes turha jatkaa tilastoaineiston matemaattista tarkastelua seuraavassa esitettävillä perusmenetelmillä. Jos taas hajontakuviossa ilmenee säännönmukaisuutta, on seuraavaksi ryhdyttävä tutkimaan tarkemmin muuttujien välistä yhteyttä. Muuttujien välistä yhteyttä voidaan tutkia matemaattisesti muun muassa siten, että lasketaan muuttujien välinen **korrelaatiokerroin** (coefficient of correlation). Korrelaatiokerroin on tilastollinen tunnusluku, jota käytetään hyväksi muiden tunnuslukujen tavoin tilastoanalyyseissä.

Korrelaatiokertoimella mitataan muuttujien välisen yhteyden voimakkuutta. Muuttujien mitta-asteikko on kuitenkin otettava huomioon, minkä vuoksi korrelaatiokertoimia on useita. Eri korrelaatiokertoimista esitellään tässä luvussa kaksi tavallisinta eli Pearsonin korrelaatiokerroin ja Spearmanin järjestyskorrelaatiokerroin.

Pearsonin korrelaatiokerroin

Yleisimmin käytetty korrelaatiokerroin on Pearsonin korrelaatiokerroin (Pearson's correlation coefficient). Usein korrelaatiokertoimesta puhuttaessa tarkoitetaankin nimenomaan Pearsonin korrelaatiokerrointa.

Tarkastellaan muuttujien X ja Y havaintoarvoista muodostettuja havaintopareja (x_1, y_1), (x_2, y_2),... (x_n, y_n), ja oletetaan, että molemmat muuttujat on mitattu **välimatka**- tai **suhdeasteikolla**. Muuttujien X ja Y välinen Pearsonin korrelaatiokerroin r määritellään seuraavasti:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum (x_i - \bar{x})^2\right] \cdot \left[\sum (y_i - \bar{y})^2\right]}}$$

Kaavassa

x_i = muuttujan X i. (luetaan: "i:nnes") havaintoarvo

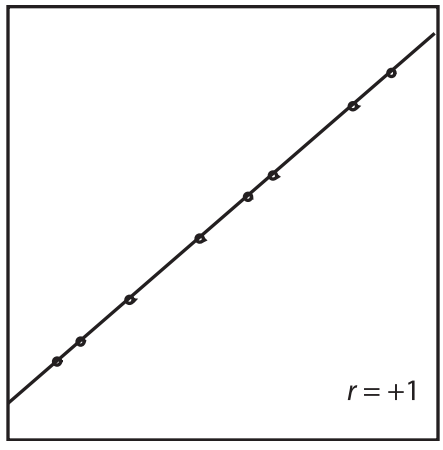
$\bar{x}$ = x_i-arvojen keskiarvo

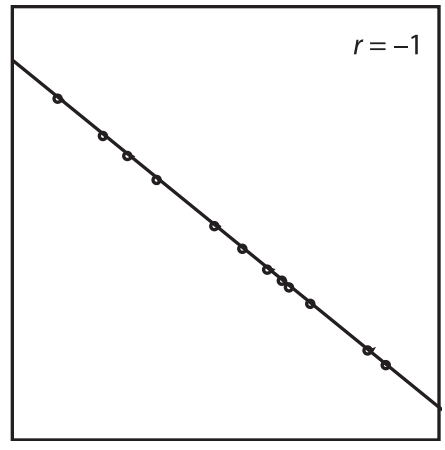
y_i = muuttujan Y i. havaintoarvo

$\bar{y}$ = y_i-arvojen keskiarvo.

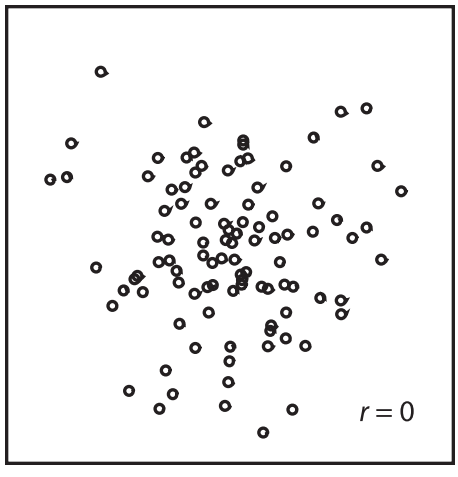
Korrelaatiokertoimen ominaisuuksia

1. Pearsonin korrelaatiokerroin r mittaa vain lineaarista yhteyttä.
2. Korrelaatiokerroin r on aina -1:n ja $+1$:n välillä oleva reaaliluku: $-1 \leq r \leq +1$.
3. Korrelaatiokertoimen arvo $+1$ saavutetaan silloin, kun kaikki hajontakuvion pisteet sijaitsevat samalla nousevalla suoralla, ja arvo -1 silloin, kun kaikki pisteet sijaitsevat samalla laskevalla suoralla. Pelkästään korrelaatiokertoimen arvon perusteella ei voida kuitenkaan päätellä, mikä on suoran kulmakerroin.





4. Mitä lähempänä korrelaatiokertoimen itseisarvo on lukua 1, sitä voimakkaampaa on muuttujien välinen lineaarinen yhteys.
5. Jos muuttujat ovat riippumattomia, muuttujien välinen korrelaatiokerroin on (ainakin likimäärin) nolla.



6. Vaikka korrelaatiokerroin olisi nolla, muuttujat eivät välttämättä ole riippumattomia. Muuttujien välinen yhteys ei ole lineaarinen, mutta muunlaista yhteyttä voi olla (katso sivun 227 kuvan c-kohtaa).
7. Korrelaatiokertoimeen ei liity mitään mittayksikköä, ja sen arvo on riippumaton muuttujien mittayksiköistä. Korrelaatiokertoimen arvo ei myöskään muutu, vaikka muuttujan *X* tai muuttujan *Y* tai molempien arvot kerrotaan (tai jaetaan) mielivaltaisella positiivisella vakiolla.
8. Korrelaatiokertoimen arvo on sama, vaikka *X*:n ja *Y*:n arvot vaihtavat paikkaa.
9. Korrelaatiokerroin on herkkä poikkeaville arvoille (outlier). Erityisesti pienissä aineistoissa yksikin toisista selvästi poikkeava arvo voi suuresti vaikuttaa korrelaatiokertoimen arvoon.

Jos hajontakuvion osoittama säännönmukaisuus on muunlaista kuin suoraviivaista, korrelaatiokertoimen arvo ei anna oikeaa kuvaa muuttujien välisen yhteyden voimakkuudesta. Esimerkiksi sivun 227 kuvan c pisteet muodostavat säännöllisen pistejoukon (paraabelin). Tästä huolimatta korrelaatiokerroin on nolla.

Korrelaatiokertoimen laskeminen

Tavallisesti korrelaatiokertoimen arvo lasketaan suoraan tilasto- tai taulukkolaskentaohjelman tai sopivan funktiolaskimen avulla. Jos tarvittavaa ohjelmaa tai laskinta ei ole käytettävissä, korrelaatiokertoimen voi laskea määritelmän mukaisella kaavalla.

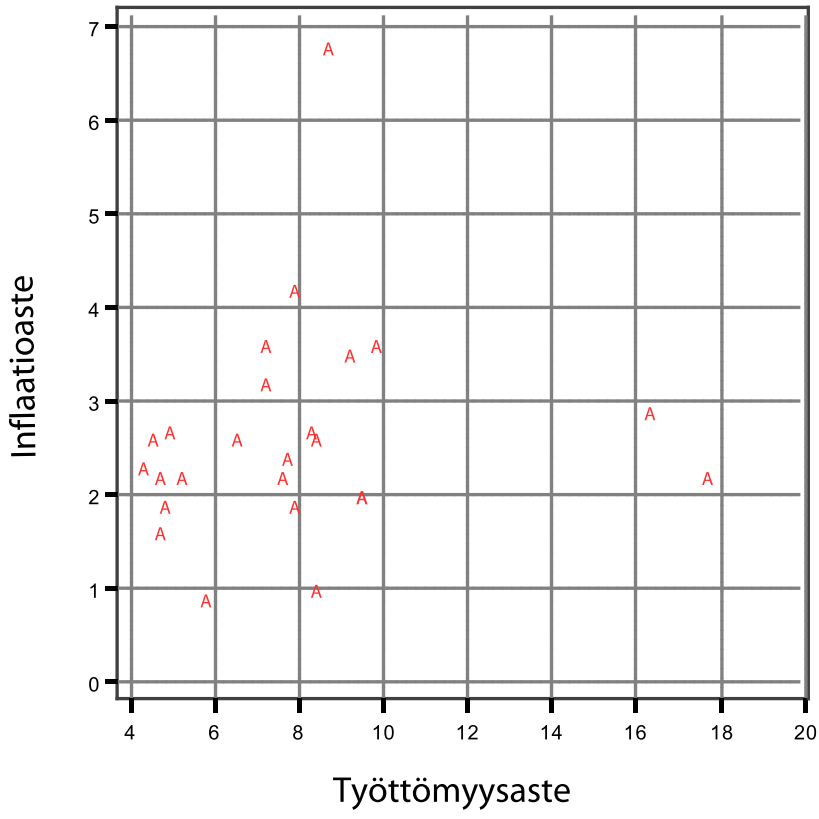
ESIMERKKI 2

Vielä 1990-luvulla eräät taloustieteilijät olivat sitä mieltä, että työttömyysaste ja inflaatioaste ovat negatiivisesti korreloivia. Myöhemmin haluttiin kuitenkin selvittää, pitikö otaksuma edelleen paikkansa. Asian selvittämiseksi tutkittiin eräiden EU-maiden työttömyysastetta ja inflaatioastetta vuodelta 2005. Tulokseksi saatiin seuraava taulukko (lähde: Tilastokeskus):

Työttömyysaste ja inflaatioaste eräissä EU-maissa

Maa	Inflaatioaste	Työttömyysaste
Alankomaat	1,5	4,7
Belgia	2,5	8,4
Iso-Britannia	2,1	4,7
Espanja	3,4	9,2
Irlanti	2,2	4,3
Italia	2,3	7,7
Itävalta	2,1	5,2
Kreikka	3,5	9,8
Kypros	2,6	4,9
Latvia	6,7	8,7
Liettua	2,6	8,3
Luxemburg	2,5	4,5
Malta	3,1	7,2
Portugali	2,1	7,6
Puola	2,1	17,7
Ranska	1,9	9,5
Ruotsi	0,8	5,8
Saksa	1,9	9,5
Slovakia	2,8	16,3
Slovenia	2,5	6,5
Suomi	0,9	8,4
Tanska	1,8	4,8
Tšekki	1,8	7,9
Unkari	3,5	7,2
Viro	4,1	7,9

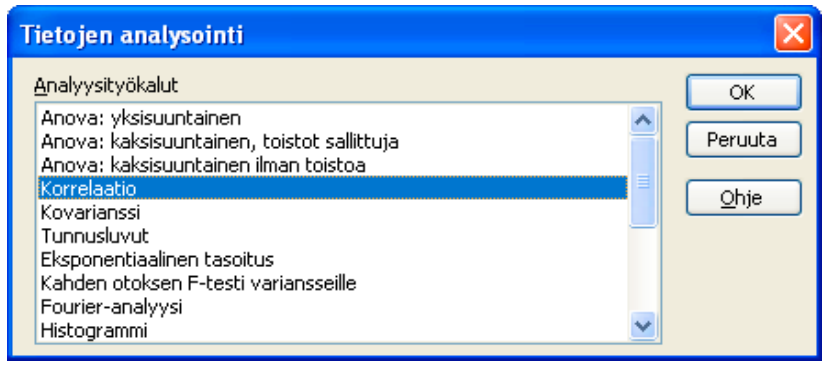
Aluksi todetaan, ettei ole olemassa mitään järkevää syytä sille, ettei työttömyysasteella ja inflaatiolla voisi olla jotakin yhteyttä toisiinsa. Niinpä muuttujien tarkastelu aloitetaan piirtämällä hajontakuvio. Tällöin saadaan alla oleva kuva.

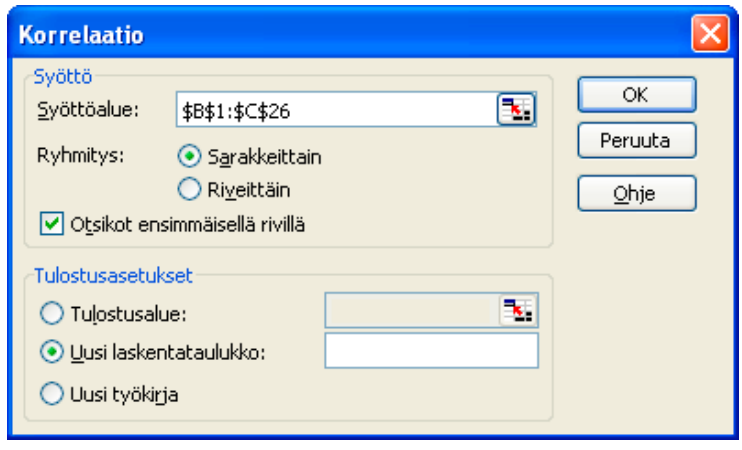


Työttömyysaste ja inflaatioaste hajontakuvioina.

Hajontakuviota tarkasteltaessa havaitaan, ettei pistejoukossa näy selvää säännönmukaisuutta, vaikka joihinkin korkeimpiin työttömyysasteisiin liittyy korkeita inflaatioasteita. Hajontakuvioon voidaan tietysti piirtää suora viiva kuvaamaan pistejoukkoa, jolloin nähtäisiin, esiintyykö pistejoukossa lineaarisuutta ja onko Pearsonin korrelaatiokertoimen laskeminen yleensäkään mielekästä.

Lasketaan korrelaatiokertoimen arvo Excelillä.





	Inflaatioaste	Työttömyysaste
Inflaatioaste	1	
Työttömyysaste	0,148047151	1

Excel antaa suoraan korrelaatiokertoimen arvoksi 0,148.

Tulkinta: Sivun 237 kuvan ja korrelaatiokertoimen arvon perusteella vaikuttaa alustavasti siltä, ettei työttömyysasteella ja inflaatioasteella ole kovin paljon tekemistä toistensa kanssa. Jos muuttujilla on jokin yhteys keskenään, se vaikuttaa pikemminkin positiiviselta kuin negatiiviselta. Johtopäätösten tekemisessä pitää kuitenkin olla varovainen. Korrelaatiokertoimen testausta tarkastellaan kohdassa 8.3 ja korrelaatiokertoimen tulkintaa kohdassa 8.4.

Korrelaatiomatriisi

Tilasto-ohjelmien avulla voidaan laskea samanaikaisesti useiden muuttujien parittaiset korrelaatiokertoimet. Tulokset esitetään tavallisesti **korrelaatiomatriisina**.

ESIMERKKI 3

Esimerkissä 1 olleiden muuttujien välinen korrelaatiomatriisi on seuraava:

	Bkt/henk. USD	**Elinajan-odote**	**Laajakais-tayhtey-det/100 henk.**	**Matkapu-helinliitty-mät/100 henk.**	**Synty-vyys/1000 henk.**
Bkt/henk. USD	1	,629	,605	,477	,590
Elinajano-dote	,629	1	,491	,099	,318
Laajakais-tayhteydet/ 100 henk.	,605	,491	1	,161	,371
Matkapu-helinliitty-mät/100 henk.	,477	,099	,161	1	0,72
Syntyvyys/ 1000 henk.	,590	,318	,371	,072	1

Korrelaatiomatriisissa päälävistäjällä on aina ykköset. Koska korrelaatiomatriisi on symmetrinen päälävistäjän suhteen, pelkkä päälävistäjän ylä- tai alapuolisen osan tulostaminen riittäisi. Tavaksi on kuitenkin tullut esittää korrelaatiomatriisi kokonaisena.

Spearmanin järjestyskorrelaatiokerroin

Pearsonin korrelaatiokerrointa käytettäessä molempien muuttujien mitta-asteikkojen on oltava vähintään välimatka-asteikkoja. Ellei tämä ehto ole voimassa, Pearsonin korrelaatiokerrointa ei voida käyttää muuttujien välisen yhteyden voimakkuuden mittaamiseen. Tällaisten

tapausten varalle on kehitetty muita muuttujien välistä yhteyttä kuvaavia mittoja. Yksi niistä on **Spearmanin järjestyskorrelaatiokerroin** (Spearman's coefficient of rank correlation), jolla on monia samoja piirteitä kuin Pearsonin korrelaatiokertoimella.

Tarkastellaan muuttujien X ja Y havaintoarvoista muodostettuja havaintopareja (x_1, y_1), (x_2, y_2),... (x_n, y_n). Oletetaan, että molemmat muuttujat on mitattu vähintään **järjestysasteikolla**. Muuttujien X ja Y välinen Spearmanin järjestyskorrelaatiokerroin r_s (Spearman's rho) määritellään seuraavasti:

$$r_S = 1 - \frac{6 \cdot \left(\sum_{i=1}^{n} d_i^{\,2} \right)}{n(n^2 - 1)}$$

d_i = i:nnen havaintoparin järjestyslukujen erotus
n = havaintoparien määrä.

Voidaan osoittaa, ettei järjestyskorrelaatiokertoimeen liity mitään mittayksikköä ja sen arvo on aina välillä [–1, 1]. Spearmanin järjestyskorrelaatiokerrointa ei kannata käyttää, jos mittaukset on tehty välimatka- tai suhdeasteikolla.

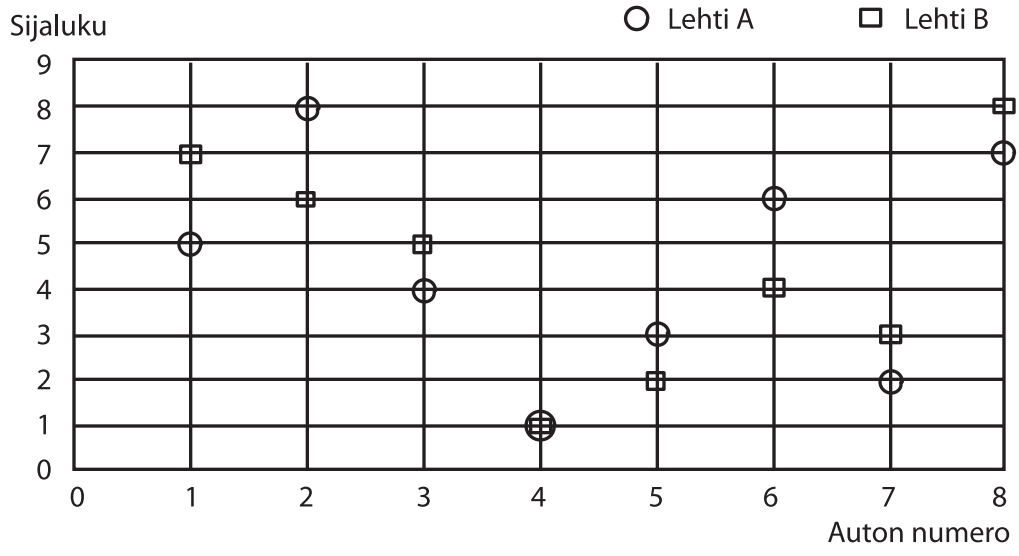
ESIMERKKI 4

Kaksi autolehteä arvioi 8 uutta keskiluokan automallia siten, että paras sai pistemäärän 1 ja huonoin pistemäärän 8. Arvioinnissa otettiin monipuolisesti huomioon auton hinta, polttoaineen kulutus, ajettavuus, huollot, jarrut, valot yms. Tulos oli seuraava:

Lehtien A ja B esittämät autojen paremmuusjärjestykset

Auto	Lehti A	Lehti B
1	5	7
2	8	6
3	4	5
4	1	1
5	3	2
6	6	4
7	2	3
8	7	8

Tutkitaan, kuinka yksimielisiä lehdet ovat autojen paremmuusjärjestyksestä. Tilannetta voidaan alustavasti tarkastella seuraavan kuvan avulla.



Autojen paremmuusjärjestyksen vertailu graafisesti.

Kuvasta käy ilmi, että mitä lähempänä pisteet ovat toisiaan, sitä yksimielisempiä lehtien arviot ovat. Näyttää siis siltä, että mielipiteillä on jonkinasteista yhtenevyyttä.

Lasketaan seuraavaksi Spearmanin järjestyskorrelaatiokerroin. Laskemisen lähtökohdaksi voidaan ottaa esimerkiksi lehden A tulos, jolloin saadaan seuraava taulukko:

Spearmanin järjestyskorrelaatiokertoimen laskeminen

Auto	Lehti A	Lehti B	d_i	d_i^2
1	5	7	–2	4
2	8	6	2	4
3	4	5	–1	1
4	1	1	0	0
5	3	2	1	1
6	6	4	2	4
7	2	3	–1	1
8	7	8	–1	1
Yhteensä				16

$$r_S = 1 - \frac{6 \cdot 16}{8 \cdot (8^2 - 1)} = 0{,}810$$

Edellä saatu järjestyskorrelaatiokertoimen arvo 0,810 näyttää osoittavan, että näkemykset autojen paremmuusjärjestyksestä ovat varsin pitkälti samansuuntaisia.

Sama tulos saadaan tilasto-ohjelman tulostamasta taulukosta:

			Lehti A	Lehti B
Spearman's rho	Lehti A	Correlation Coefficient	1,000	,810
		Sig. (2-tailed)	.	,015
		N	8	8
	Lehti B	Correlation Coefficient	,810*	1,000
		Sig. (2-tailed)	,015	.
		N	8	8

8.3 Korrelaatiokertoimen merkitsevyys

Korrelaatiokertoimen arvo poikkeaa yleensä nollasta. Peruskysymys on, milloin poikkeama nollasta on niin suuri, ettei sitä voida katsoa pelkästään satunnaisista syistä johtuvaksi. Tämän selvittämiseksi korrelaatiokertoimen arvo yleensä testataan. Useisiin tilasto-ohjelmiin kuuluu korrelaatiokertoimen testaus, jossa annettua riskitasoa ja otoskokoa vastaava kriittinen arvo saadaan tulostetuksi tai ohjelma antaa suoraan *p*-arvon. *P*-arvo on erehtymisriski, kun nollahypoteesi hylätään.

Pearsonin korrelaatiokertoimen testauksen ja myös kirjan liitteenä olevan taulukon käyttö edellyttää, että

- molemmat muuttujat noudattavat ainakin likimäärin normaalijakaumaa (ei välttämättä samaa)
- tapaukset muodostavat satunnaisotoksen perusjoukosta
- muuttujien saamat arvot yhden tapauksen osalta ovat riippumattomia muuttujien saamista arvoista muiden tapausten osalta (esimerkiksi sivun 231 taulukossa eri maiden tulokset muuttujalle Bkt/henkilö ovat riippumattomia muiden maiden tilanteesta).

Jos hajontakuvio antaa aihetta epäillä, ettei muuttujien välinen yhteys ole suoraviivainen, korrelaatiokertoimen testaus ei välttämättä anna totuudenmukaista kuvaa muuttujien välisestä yhteydestä.

Korrelaatiokertoimen testaamisessa testin nollahypoteesina on se, että Pearsonin korrelaatiokerroin on nolla ($r = 0$). Tämä vastaa tilannetta, jossa muuttujat ovat toisistaan lineaarisesti riippumattomia. Testisuureena käytetään lauseketta

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

minkä jälkeen testisuureen saamaa arvoa verrataan *t*-jakauman taulukon arvoon, kun vapausasteita (degrees of freedom) *f (tai df)* on $n - 2$. Lopuksi tehdään johtopäätös.

ESIMERKKI 5

Tarkastellaan esimerkissä 2 esitettyä otaksumaa, että työttömyysasteen ja inflaatioasteen välillä vallitsisi negatiivinen korrelaatio. Tätä varten lasketaan testisuureen arvo

$$t = \frac{0{,}148 \cdot \sqrt{25-2}}{\sqrt{1-0{,}148^2}} = 0{,}718.$$

Jos riskitasoksi valitaan 5 %, *t*-jakauman taulukosta saadaan kaksisuuntaisen testin kriittiseksi arvoksi 2,069. Koska laskettu testisuureen itseisarvo on alle kriittisen arvon, toisin sanottuna 0,718 < 2,069, edellä olevan perusteella ei voida osoittaa, että työttömyysasteen ja inflaatioasteen välillä vallitsisi positiivinen tai negatiivinen korrelaatio. Johtopäätöksen erehtymisriski on alle 5 %. Se, että esimerkissä yhteyttä ei löytynyt, voi johtua siitäkin, että jotkin EU:n direktiivit säätelevät EU-maiden inflaatioastetta. Sama johtopäätös voidaan tehdä suoraan tilasto-ohjelman tulostamasta taulukosta:

		Inflaatioaste	Työttömyysaste
Inflaatioaste	Pearson Correlation	1	,148
	Sig. (2-tailed)		,480
	N	25	25
Työttömyysaste	Pearson Correlation	,148	1
	Sig. (2-tailed)	,480	
	N	25	25

Taulukon *p*-arvo lasketaan nollahypoteesille: muuttujien välinen korrelaatiokerroin on nolla. Koska nollahypoteesin hylkäämiseen liittyvän riskin todennäköisyys on taulukon mukaan 0,480 ja se on selvästi suurempi kuin yleisesti käytetty 5 %:n riskitaso, muuttujien välillä ei tämän aineiston perusteella voida osoittaa olevan tilastollista yhteyttä.

Korrelaatiokerroin voidaan suoraan testata myös valmiiden taulukoiden avulla, jolloin vältytään testisuureen laskemiselta. Kirjan liitteenä olevan taulukon avulla voidaan testata korrelaatiokertoimen merkitsevyys. Esimerkin 5 testaaminen taulukkoa käyttäen tapahtuu seuraavasti. Otoskokoa $n = 25$ vastaava korrelaatiokertoimen kriittinen arvo on 0,396, kun merkitsevyystaso on kaksisuuntaisessa testissä 5 %. Koska laskettu korrelaatiokertoimen itseisarvo on pienempi kuin taulukossa oleva kriittinen arvo ($0{,}148 < 0{,}396$), edellä olevan perusteella ei voida osoittaa, että työttömyysasteen ja inflaatioasteen välillä vallitsisi positiivinen tai negatiivinen korrelaatio.

ESIMERKKI 6

Tarkastellaan aikaisemmin esimerkissä 3 laskettua korrelaatiomatriisia. Jos kaksisuuntaisen testin merkitsevyystasoksi valitaan nyt 1 % ja otoskoko on 25, korrelaatiokertoimen kriittinen arvo on kirjan liitteenä olevan taulukon mukaan 0,505. Jos siis korrelaatiokertoimen itseisarvo on pienempi kuin 0,505, muuttujien välillä ei voida osoittaa mitään tilastollista yhteyttä. Oheisen taulukon mukaan näyttää siltä, että bkt:llä henkilöä kohden on yhteys elinajanodotteeseen, laajakaistayhteyksien lukumäärään ja syntyvyyteen. Muita yhteyksiä ei 1 %:n merkitsevyystasolla voida tämän aineiston perusteella osoittaa.

	Bkt/henk. USD	**Elinajan-odote**	**Laajakais-tayhtey-det/100 henk.**	**Matkapu-helinliitty-mät/100 henk.**	**Synty-vyys/1000 henk.**
Bkt/henk. USD	1	**,629**	**,605**	,477	**,590**
Elinajan-odote	**,629**	1	,491	,099	,318
Laajakais-tayhteydet/ 100 henk.	**,605**	,491	1	,161	,371
Matkapu-helinliitty-mät/100 henk.	,477	,099	,161	1	0,72
Syntyvyys/ 1000 henk.	**,590**	,318	,371	,072	1

Järjestyskorrelaatiokertoimen testaamiseen käytetään samaa testisuuretta kuin Pearsonin korrelaatiokertoimen testaamiseen:

$$t = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^{\,2}}}.$$

Testisuureen saamaa arvoa verrataan sitten *t*-jakauman taulukon arvoon, kun vapausasteita *f* on *n* – 2. Lopuksi tehdään johtopäätös.

ESIMERKKI 7

Testataan 5 %:n merkitsevyystasoa käyttämällä, poikkeaako esimerkissä 4 laskettu Spearmanin järjestyskorrelaatiokerroin nollasta. Tätä varten lasketaan testisuureen *t* arvo:

$$t = \frac{0{,}810 \cdot \sqrt{8-2}}{\sqrt{1-0{,}810^{\,2}}} = 3{,}383.$$

Kun merkitsevyystaso on 5 % ja havaintoparien lukumäärä 8, kaksisuuntaisen *t*-testin kriittinen arvo on 2,447. Koska testisuureen laskettu arvo on suurempi kuin taulukosta löytyvä kriittinen arvo, toisin sanottuna 3,383 > 2,447, Spearmanin järjestyskorrelaatiokertoimen korkea arvo ei ole pelkästään sattuman tulosta. Lehtien näkemykset autojen paremmuusjärjestyksestä ovat siten yhdensuuntaiset. Väärän johtopäätöksen riski on alle 5 %.

Joskus korrelaatiokertoimen testauksen yhteydessä saadaan tulos, jonka mukaan muuttujien välillä ei voida osoittaa tilastollista yhteyttä, vaikka testaajalla on asiasta vahva ennakkokäsitys. Tällöin on tehtävä uusi tutkimus, mielellään edellistä suuremmalla aineistolla, ja toivottava, että oletettu yhteys ilmestyisi tilastollisesti näkyviin.

8.4 Kertoimien tulkinta

Jos Pearsonin korrelaatiokertoimen arvo on positiivinen, muuttujien arvot muuttuvat samaan suuntaan. Tällöin *X*:n arvojen kasvaessa myös *Y*:n arvot yleisesti kasvavat. Jos korrelaatiokertoimen arvo on lähellä ykköstä, muuttujien välillä on voimakas positiivinen lineaarinen yhteys. Jos taas korrelaatiokertoimen arvo on negatiivinen, muuttujien *X* ja *Y* arvot muuttuvat vastakkaisiin suuntiin eli *X*:n arvojen kasvaessa

Y:n arvot pienenevät. Jos korrelaatiokertoimen arvo on lähellä lukua –1, muuttujien välillä on voimakas negatiivinen lineaarinen yhteys. Korrelaatiokertoimen etumerkki osoittaa siis muuttujien välisen yhteyden suunnan.

Jos Pearsonin korrelaatiokertoimen arvo on lähellä nollaa, muuttujien välillä ei ole lineaarista yhteyttä. Vaikka korrelaatiokertoimen arvo poikkeaa nollasta ja on jo lähellä lukua 0,5 tai –0,5, se osoittaa vielä melko heikkoa lineaarista yhteyttä muuttujien välillä. Huomattakoon kuitenkin, että heikkokin muuttujien välinen yhteys saattaa olla käytännössä tärkeä. Esimerkiksi yhteiskunnallisissa tutkimuksissa korrelaatiot jäävät usein melko pieniksi mutta ne ovat silti tärkeitä.

Korrelaatiokertoimen tulkinnassa pitää olla varovainen. On väitetty, että korrelaatiokerroin on se tilastollinen tunnusluku, jota käytetään kaikista eniten väärin. Korrelaatiokertoimen suurikaan itseisarvo ei takaa muuttujien välillä vallitsevaa syy-seuraussuhdetta. Esimerkiksi eräänä vuonna havaittu korkea positiivinen korrelaatio uima-asujen myynnin ja jäätelön myynnin välillä ei oikeuttanut johtopäätökseen, että uima-asujen myynnillä olisi ollut vaikutusta jäätelön menekkiin. Mahdollisesti molempien artikkeleiden myyntiä vauhditti lämmin kesä!

Toisaalta edes se, että korrelaatiokertoimen arvo on nolla tai lähellä nollaa, ei tarkoita sitä, ettei muuttujien välillä olisi mitään yhteyttä. Yhteys ei vain ole silloin lineaarista. Muunlaista yhteyttä voi kyllä esiintyä. Hajontakuvio osoittautuu tässäkin tapauksessa tärkeäksi.

Erityisesti pienissä aineistoissa korrelaatiokertoimen itseisarvon suuri arvo aiheuttaa toisinaan virhetulkintoja. Jos merkitsevyystasoksi valitaan 5 % eli 0,05, sadassa toisilleen riippumattomien muuttujien välisen korrelaatiokertoimen arvoista löytyy keskimäärin viisi lukua, jotka ylittävät korrelaatiokertoimen kriittisen arvon. Tämä johtuu siis pelkästään sattumasta ilman, että muuttujien välillä todellakaan tarvitsisi olla mitään yhteyttä keskenään.

Spearmanin järjestyskorrelaatiokerroin on käyttökelpoinen esimerkiksi silloin, kun on selvitettävä, ovatko arvioitsijoiden mielipiteet yhdensuuntaiset vai erisuuntaiset. Jos järjestyskorrelaatiokerroin on positiivinen, arvioitsijoiden mielipiteet ovat samansuuntaisia. Negatiivinen järjestyskorrelaatiokertoimen arvo kertoo erimielisyydestä. Jos järjestyskorrelaatiokerroin on lähellä nollaa, järjestyksestä ei ole erimielisyyttä muttei myöskään yksimielisyyttä.

Tulkintavirheitä

Korrelaatiokertoimen tulkintavirheitä aiheuttavat useimmiten seuraavat seikat:

– Riippuvuudesta ei välttämättä seuraa syy-seurausyhteyttä.
– Kolmas muuttuja eli kahden muuttujan välinen yhteys selittyy yhteisestä syystä (esimerkiksi lämpimästä kesästä).
– Muuttujien välinen yhteys ei ole lineaarinen.
– Havainto on poikkeava, sillä korrelaatiokerroin on herkkä poikkeaville arvoille (ks. kohtaa 9.5).

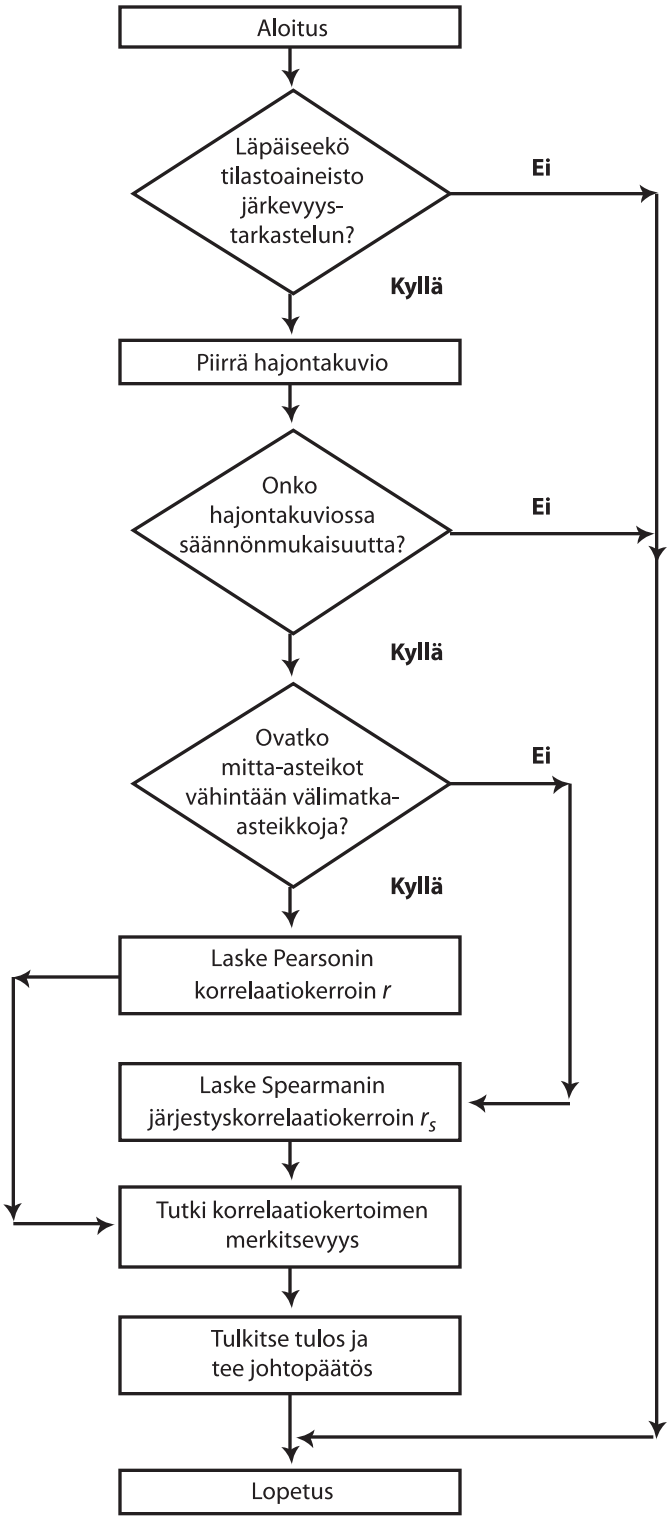
Puutteita

Korrelaatiokertoimella on kaksi puutetta:

– Se mittaa vain lineaarista riippuvuutta.
– Se ei ole malli, jonka avulla nähtäisiin, miten toinen muuttuja vaikuttaa toiseen muuttujaan.

8.5 Yhteyden tutkimiseen liittyvä vuokaavio

Tilastollisen yhteyden tutkiminen voidaan esittää yksinkertaistetulla vuokaaviolla.



8.6 Case B: Pakokaasujen pitoisuudet

Tutkimuksen tausta

Bensiinikäyttöisistä ajoneuvoista haluttiin tutkia haitallisten pakokaasujen määriä. Asian selvittämiseksi mitattiin katsastuksen yhteydessä ilman katalysaattoria olevien ajoneuvojen häkä- ja hiilivetypitoisuudet. Häkä eli CO ilmoitetaan pakokaasun prosentuaalisena osuutena. Hiilivedyn eli CH:n yksikkönä on ppm eli miljoonasosa.

Haitallisten pakokaasujen enimmäismäärät joutokäynnillä

Ajoneuvon käyttöönotto	CO (%) (häkä)	CH (ppm) (hiilivety)
1.1.1986 tai sen jälkeen	3,5	600
Katalysaattoriajoneuvo	0,3	100

On arveltu, että vain toisen haitallisen pakokaasun määrän mittaaminen riittäisi. Jos suuria häkäpitoisuuksia esiintyisi säännönmukaisesti suurien hiilivetypitoisuuksien yhteydessä, säästettäisiin sekä aikaa että laitekustannuksia mittaamalla vain toinen näistä pitoisuuksista. Arvelujen todentamiseksi katsastusasemalla kirjattiin 20 ajoneuvon pakokaasujen pitoisuudet.

Pakokaasujen mittaustulokset

CH	CO	CH	CO	CH	CO	CH	CO
189	0,23	180	0,25	307	2,72	352	1,27
151	1,17	187	2,52	290	3,30	229	0,22
410	2,38	294	1,75	203	2,33	210	1,80
486	6,31	288	5,93	300	2,24	333	1,93
215	0,51	129	4,06	513	3,14	259	0,60

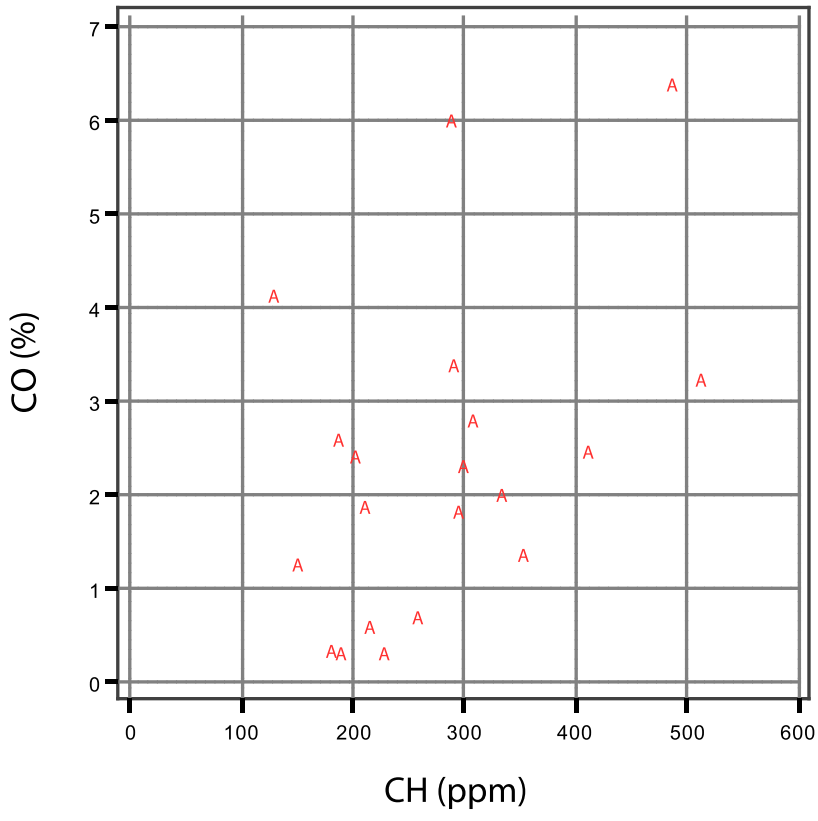
Järkevyystarkastelu

Järkevyystarkastelun aluksi harkitaan, voivatko hiilivedyn ja hään eli hiilimonoksidin pitoisuudet olla missään yhteydessä toisiinsa. Koska molemmat pakokaasut ovat hiiltä sisältäviä polttoaineen palamistuloksia, ei ole ainakaan mitään loogista estettä sille, etteivät pakokaasujen pitoisuudet voisi olla jossakin yhteydessä toisiinsa. Etukäteen on

kuitenkin vaikea sanoa, että toinen näistä pakokaasuista olisi toisen esiintymisen syy tai seuraus. Niinpä kumpaakin muuttujaa voidaan pitää satunnaismuuttujana.

Hajontakuvio

Seuraavaksi piirretään tilastoaineistoa esittävä hajontakuvio. Alla olevasta kuviosta käy ilmi, että pistejoukko on melko hajallaan, joskin lievää positiivista yhteyttä muuttujilla saattaa kuvion mukaan olla.



Pakokaasujen hajontakuvio.

Korrelaatiokerroin

Seuraavaksi tutkitaan muuttujien välisen yhteyden voimakkuus. Koska mitta-asteikot eivät aseta mitään rajoituksia, valitaan mitaksi Pearsonin korrelaatiokerroin. Tulokset on esitetty seuraavassa taulukossa.

Hiilivedyn ja hiilimonoksidin välinen korrelaatio

		CH (ppm)	CO (%)
CH (ppm)	Pearson Correlation	1	,445
	Sig. (2-tailed)		,049
	N	20	20
CO (%)	Pearson Correlation	,445	1
	Sig. (2-tailed)	,049	
	N	20	20

Taulukosta käy ilmi, että muuttujien välillä näyttää olevan lievä positiivinen korrelaatio. On kuitenkin vielä testattava, onko tulos mahdollisesti vain sattuman aiheuttamaa vai osoittaako se selkeästi muuttujien välisen yhteyden olemassaolon. Asian selvittämiseksi tutkitaan liitteenä olevan taulukon avulla, mikä on korrelaatiokertoimen kriittinen arvo, kun otoskoko on 20:

– 5 %:n riskitasolla kriittinen arvo on 0,444
– 1 %:n riskitasolla kriittinen arvo on 0,561.

Tulkinta

Laskettu korrelaatiokertoimen arvo 0,445 ylittää hiuksenhienosti 5 %:n kriittisen arvon mutta on selvästi 1 %:n kriittisen rajan alapuolella. Tulos antaa siis aiheen epäillä, että suuret häkä- ja hiilivetypitoisuudet mahdollisesti esiintyvät samanaikaisesti. Erehtymisriski on kuitenkin melko suuri erityisesti siksi, että tässä käytetty tilastoaineisto on kovin pieni. Jos siis uskotaan, että hiilivedyn ja hiilimonoksidin pitoisuuksien välillä todella vallitsee yhteys, on poimittava uusi entistä suurempi otos.

Pakokaasujen pitoisuuksissa lienee myös sellaisia eroja, jotka riippuvat ajoneuvon iästä, merkistä, mallista, ajokilometrien määrästä ja huollon ajankohdasta. Jos halutaan, että nämä seikat näkyvät tuloksessa, otoskokoa on suurennettava reilusti. Mainittakoon myös, ettei hiilivety- ja häkäpitoisuuksien välillä ole viranomaisten mielestä yhteyttä, joten ne mitataan aina erikseen.

TEHTÄVIÄ

1. Onko totta, että
   a) jos kahdella tilastoaineistolla on sama korrelaatiokerroin, aineistot ovat identtiset
   b) jos muuttujien *x* ja *y* välinen korrelaatiokerroin on positiivinen, *x*:n arvojen pienentyessä myös *y*:n arvot pienenevät?

2. Onko totta, että vaikka yhden muuttujan euroina ilmoitettu rahamäärä muutetaan dollareiksi, vastaava korrelaatiokerroin pysyy kuitenkin samana?

3. Onko totta, että jos hajontakuvion pisteet näyttävät noudattavan paraabelin muotoa, korrelaatiokerrointa ei voida laskea?

4. Mitkä ovat ne erikoistapaukset, joissa Pearsonin korrelaatiokerroin ei ole +1 eikä –1, vaikka kaikki pisteet sijaitsevat samalla suoralla?

5. Mikä johtopäätös yritysjohdon pitää tehdä, jos myynti on riippumatonta mainonnasta?

6. Osoittaako korrelaatiokertoimen arvo +0,35 voimakkaampaa yhteyttä muuttujien välillä kuin korrelaatiokertoimen arvo –0,80?

7. Haluttiin tutkia, onko bruttokansantulolla (ostovoimakorjattu USD/henkilö) ja elinajanodotteella yhteyttä. Tätä varten kerättiin oheinen tilasto eräistä EU-maista vuodelta 2006 (lähde: Tilastokeskus).

Maa	Bkt/henk. (USD)	Elinajanodote
Alankomaat	40 552	79,0
Belgia	37 164	78,8
Iso-Britannia	38 947	78,5
Espanja	29 266	79,7
Irlanti	52 360	77,7
Italia	31 350	79,8
Itävalta	38 865	79,1
Kreikka	21 925	79,2
Kypros	21 088	77,8
Latvia	8 349	71,3
Liettua	8 419	74,2
Luxemburg	85 444	78,9
Malta	14 053	79,0
Portugali	18 161	77,7
Puola	8 655	75,0
Ranska	35 336	79,7
Ruotsi	41 945	80,5
Saksa	35 022	78,8
Slovakia	9 808	74,7
Slovenia	17 267	76,3
Suomi	39 332	78,5
Tanska	50 808	77,8
Tšekki	13 654	76,2
Unkari	10 298	72,7
Viro	11 238	72,0

a) Piirrä tilastoaineistoa kuvaava hajontakuvio.
b) Määritä Pearsonin korrelaatiokerroin.
c) Vertaa saatua tulosta esimerkin 3 taulukkoon.

8. Haluttiin tutkia avioliittojen solmimisen ja syntyvyyden välistä yhteyttä. Asian selvittämiseksi kerättiin oheinen tilasto eräistä EU-maista vuodelta 2005 (lähde: Tilastokeskus).

Maa	Avioliitot/ 1000 henk.	Syntyvyys/ 1000 henk.
Alankomaat	4,5	11,5
Belgia	4,1	11,2
Iso-Britannia	5,2	12,0
Espanja	4,8	10,7
Irlanti	5,0	14,7
Italia	4,3	9,5
Itävalta	4,8	9,5
Kreikka	5,5	9,7
Kypros	7,8	10,9
Latvia	5,5	9,3
Liettua	5,8	8,9
Luxemburg	4,4	11,7
Malta	5,9	9,6
Portugali	4,6	10,4
Puola	5,4	9,5
Ranska	4,5	12,9
Ruotsi	4,9	11,2
Saksa	4,7	8,3
Slovakia	4,9	10,1
Slovenia	2,9	9,1
Suomi	5,6	11,0
Tanska	6,7	11,9
Tšekki	5,1	10,0
Unkari	4,4	9,7
Viro	4,6	10,7

a) Piirrä tilastoaineistosta hajontakuvio.
b) Määritä Pearsonin korrelaatiokerroin.

9. Tutki, onko imeväiskuolleisuus 1 000 elävänä syntynyttä lasta kohden yhteydessä henkilöiden lukumäärään lääkäriä kohden. Oheinen tilasto on imeväiskuolleisuuden osalta vuodelta 2005. Henkilöiden lukumäärä lääkäriä kohden joidenkin maiden osalta on muutamaa vuotta vanhempi. (Lähde: Tilastokeskus.)

Maa	Imeväiskuolleisuus/1000	Henkilöä/lääkäri
Alankomaat	4,9	307
Belgia	4,4	223
Iso-Britannia	5,1	826
Espanja	4,1	223
Irlanti	4,0	389
Italia	4,7	162
Itävalta	4,2	216
Kreikka	3,8	221
Kypros	4,6	390
Latvia	7,8	295
Liettua	6,8	252
Luxemburg	2,6	367
Malta	6,0	344
Portugali	3,5	310
Puola	6,4	446
Ranska	3,8	306
Ruotsi	2,4	352
Saksa	3,9	271
Slovakia	7,2	263
Slovenia	4,1	430
Suomi	3,0	313
Tanska	4,4	365
Tšekki	3,4	230
Unkari	6,2	260
Viro	5,4	316

a) Piirrä tilastoaineistosta hajontakuvio.
b) Määritä Pearsonin korrelaatiokerroin.

10. Haluttiin tutkia pesukoneiden hinnan ja suurimman linkousnopeuden välistä yhteyttä. Asian selvittämiseksi kerättiin eri valmistajien esitteistä tarvittavat tilastotiedot. Oheisessa taulukossa on esitteiden tietoja.

Merkki	Hinta (€)	Linkousnopeus (k/min)
A	974	1200
B	769	1300
C	739	1300
D	671	1200
E	671	1000
F	655	1200
G	587	1200
H	503	800
I	470	800
J	756	1400
K	604	1000

a) Piirrä hintaa ja linkousnopeutta esittävä hajontakuvio.
b) Määritä hinnan ja linkousnopeuden välinen Pearsonin korrelaatiokerroin.

11. Avaa puhelinluettelo satunnaisesta kohdasta ja muodosta puhelinnumeroiden kahdesta viimeisestä numerosta muuttujien arvoja seuraavasti. Kymmenestä ensimmäisestä puhelinnumerosta muodostetaan muuttujan X arvot ja kymmenestä seuraavasta puhelinnumerosta muodostetaan muuttujan Y arvot.
a) Piirrä hajontakuvio.
b) Määritä Pearsonin korrelaatiokerroin.

12. Haluttiin tutkia säiliöpakastimien hinnan, pakastustehon ja bruttotilavuuden välistä yhteyttä. Asian selvittämiseksi kerättiin eri valmistajien esitteistä säiliöpakastimien ohjehintoja, valmistajan ilmoittamia pakastustehoja ja säiliöpakastimien bruttotilavuutta.

Tulos oli seuraava:

Säiliöpakastin nro	Hinta (€)	Pakastusteho (kg/vrk)	Bruttotilavuus (l)
1	453	26	261
2	385	23	181
3	385	20	189
4	335	20	163
5	234	7	105
6	436	24	257
7	409	21	235
8	480	30	282
9	285	15	144

a) Piirrä samanaikaiset hajontakuviot tilastoaineistosta.
b) Laske kaikki mahdolliset korrelaatiokertoimet.

13. Testaa sekä testisuuretta *t* että kirjan liitteenä olevaa taulukkoa käyttämällä tehtävien 7–11 korrelaatiokerrointen merkitsevyys 5 %:n merkitsevyystasolla.

14. Testaa kirjan liitteenä olevaa taulukkoa käyttämällä, mikä on esimerkin 4 tulos, kun merkitsevyystasoksi valitaan 5 %.

15. Kaksi henkilöä asettaa kahdeksan presidenttiehdokasta (A–H) äänestysjärjestykseen seuraavasti:

Ehdokas	A	B	C	D	E	F	G	H
Henkilö 1	1	6	5	2	4	3	7	8
Henkilö 2	2	8	6	1	4	3	5	7

Tutki, ovatko mielipiteet samansuuntaisia määrittämällä Spearmanin järjestyskorrelaatiokerroin. Testaa ja tulkitse tulos. Käytä 5 %:n merkitsevyystasoa.

16. Työntekijä- ja työnantajaosapuoli laativat listan seitsemän käsiteltävän asian tärkeysjärjestyksestä. Tulos oli seuraava:

Asia	A	B	C	D	E	F	G
Työntekijät	3	5	2	7	1	4	6
Työnantajat	2	5	7	3	6	4	1

Määritä Spearmanin järjestyskorrelaatiokerroin. Testaa ja tulkitse tulos. Käytä 5 %:n merkitsevyystasoa. Määritä myös tilastoaineiston Pearsonin korrelaatiokerroin. Vertaa tulosta Spearmanin korrelaatiokertoimeen.

17. Työpaikkahaastatteluun valittiin 10 hakijaa. Kaksi henkilöä tutki haastateltavat toisistaan riippumattomasti ja asetti heidät paremmuusjärjestykseen. Toinen henkilö tutki muodollisen pätevyyden, koulutuksen, työkokemuksen ym. taustan ja toinen arvioi haastateltavan antamaa vaikutelmaa. Tulokset olivat seuraavat:

Henkilö	A	B	C	D	E	F	G	H	I	J
Tausta	3	4	8	6	5	2	7	9	10	1
Haastattelu	4	3	5	8	7	1	6	10	9	2

Testaa sekä testisuureen *t* että kirjan liitteenä olevan taulukon avulla, onko näyttöä siitä, että tässä tapauksessa taustatiedot ja haastatteluvaikutelma ovat yhtenevät. Käytä 1 %:n merkitsevyystasoa.

18. Tutkittiin valaistuksen merkitystä tarkkuutta vaativaan työsuoritukseen. Asian selvittämiseksi valittiin 10 vapaaehtoista, jotka pujottivat langan viisi kertaa silmäneulaan yksin huoneessa, jossa valaistus oli heikko ja taustana oli musta seinä. Toisella koekerralla valaistus oli voimakas ja taustana oli valkoinen seinä. Koetulosten yhteisajat olivat seuraavat:

Henkilö	A	B	C	D	E	F	G	H	I	J
Heikko valaistus	9,2	10,4	12,8	8,6	15,9	14,7	13,5	11,8	17,7	13,3
Voimakas valaistus	9,0	9,5	10,3	8,4	16,2	9,6	11,9	10,0	14,4	12,1

a) Laske aikojen välinen Pearsonin korrelaatiokerroin ja testaa sen merkitsevyys. Käytä 1 %:n merkitsevyystasoa.
b) Laske aikojen välinen Spearmanin järjestyskorrelaatiokerroin ja testaa sen merkitsevyys. Käytä 1 %:n merkitsevyystasoa.
c) Vertaa a)- ja b)-kohdan tuloksia keskenään.

9
Regressio

Edellisessä luvussa tarkasteltiin tilastollista riippuvuutta. Kun tilastollinen riippuvuus on havaittu, ryhdytään selvittämään, onko muuttujien välillä myös syy-yhteyttä. Pelkkä riippuvuus ei nimittäin vielä takaa sitä, että toinen tekijä aiheuttaa toisen tai vaikuttaa toiseen.

Ikuinen elämä on pitkään kiinnostanut ihmistä, ja eliniän pidentämiseksi on yritetty erilaisia keinoja, kuten terveellistä ruokavaliota, säännöllistä liikuntaa, nautintoaineiden kohtuukäyttöä ja säännöllisiä elämäntapoja. Elinikään vaikuttavat varmasti myös monet muut seikat, kuten perimä ja elintaso. Tässä luvussa yritetään selvittää joitakin niitä tekijöitä, jotka voivat vaikuttaa elinajanodotteeseen. Samalla jokainen voi pohtia, mihin tekijöihin voi itse vaikuttaa ja mihin ei.

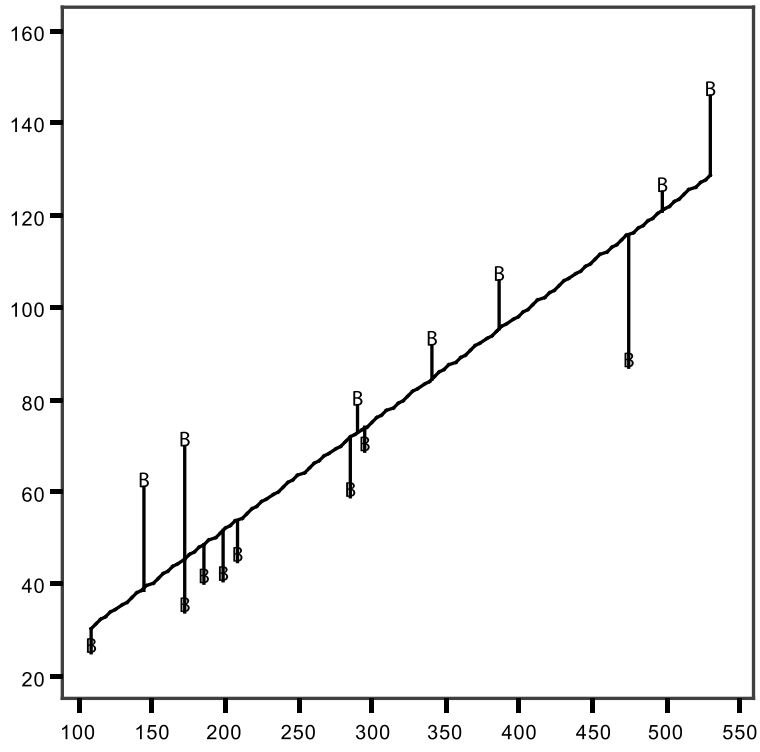
9.1 Pienimmän neliösumman menetelmä

Kun tutkitaan muuttujien välistä yhteyttä, joudutaan käymään läpi samat vaiheet, joita käsiteltiin korrelaatiokertoimen määrittämisen yhteydessä. Aluksi joudutaan siis selvittämään, voiko muuttujilla olla mitään yhteyttä keskenään. Jos tilastoaineisto läpäisee järkevyystarkastelun, piirretään aineistoa kuvaava hajontakuvio, minkä jälkeen tutkitaan saatua pistejoukkoa ja yritetään löytää siitä säännönmukaisuutta.

Jos tilastoaineistossa on havaittavissa säännönmukaisuutta, tutkitaan, voidaanko sitä kuvata jollakin tavalla, jolloin päästään **malliajatteluun**. Tehtävänä on siis löytää sellainen matemaattinen malli, että se jossakin mielessä parhaiten kuvaa saatua pistejoukkoa. Malli on sitä parempi, mitä läheisemmin se kuvaa koko pistejoukkoa. Pienimmän neliösumman menetelmä johtaa usein hyvään tulokseen. Tästä menetelmästä käytetään myös lyhennettä OLS (ordinary least squares).

Pienimmän neliösumman menetelmään tutustutaan seuraavan kuvan avulla. Siinä pistejoukkoa kuvataan suoralla, joka kulkee pistejou-

kossa mahdollisimman luontevasti. Tällöin suoran ylä- ja alapuolelle jää suunnilleen yhtä monta pistettä. Jokaisesta havaintoarvoa vastaavasta pisteestä (x_i, y_i) lasketaan y-akselin suuntainen etäisyys mallina olevasta suorasta.



Pienimmän neliösumman menetelmän mukaisen suoran määrittäminen.

Merkitään x_i:n arvolla laskettua suoran pistettä $(x_i, \hat{y}_i)$:llä. Nimensä mukaisesti **pienimmän neliösumman menetelmässä** haetaan ratkaisua lausekkeelle

$$Min \sum (y_i - \hat{y}_i)^2 = Min \sum e_i^2$$

Minimi löydetään matemaattisesti derivoimalla, sillä kyseessä on ääriarvotehtävä. Tehtävä on matemaattisesti helpoin ratkaista silloin, kun malliksi valitaan suora. Tietokoneohjelmien avulla voidaan laajentaa mallivalikoimaa.

Kun pistejoukkoa pyritään kuvaamaan matemaattisen mallin avulla, menettelyä kutsutaan **käyrän sovitukseksi** (curve fitting)**.** Tällöin on ratkaistava neljänlaisia ongelmia seuraavassa järjestyksessä:

1. Päätetään, mitkä muuttujat valitaan selittäviksi muuttujiksi.
2. Ratkaistaan matemaattisen mallin tyyppi eli se, minkälainen funktio on muodoltaan.
3. Etsitään mallin vakioiden eli parametrien arvot, jotta malli olisi täysin määrätty.
4. Tutkitaan mallin hyvyyttä, jotta saataisiin selville kuinka hyvin malli kuvaa havaintoaineistoa. Mallin hyvyyttä käytetään sitten osaltaan hyväksi ennusteiden luotettavuuden arvioimisessa.

9.2 Lineaarinen malli

Tilastollisten tarkastelujen yhtenä tavoitteena on pyrkiä kuvaamaan ilmiöitä ja löytämään niille selityksiä. Ideaalitapauksessa (esimerkiksi fysiikassa) löydetään syy-seuraussuhde siten, että jos tietyt ehdot ovat voimassa, ne aiheuttavat tiettyjä seuraamuksia. **Regressioanalyysin** päämääränä on löytää muuttujien välillä mahdollisesti vallitseva yhteys ja kuvata sitä matemaattisen mallin avulla.

Jos muuttujia on vain kaksi, toista muuttujaa kutsutaan **selittäväksi** muuttujaksi (regressor) ja toista **selitettäväksi** muuttujaksi (regressand). Selitettävää muuttujaa merkitään y:llä ja selittävää muuttujaa x:llä. Selittäviä muuttujia voi olla useampiakin kuin yksi, jolloin niistä käytetään merkintöjä $x_1, x_2, \ldots x_k$. Kohdassa 9.4 tarkastellaan tilanteita, joissa selittäviä muuttujia on useita.

Yksinkertaisimmassa **lineaarisessa mallissa** tarkasteltavaa ilmiötä pyritään kuvaamaan yhden selittävän muuttujan avulla. Yksinkertaisin kaikista matemaattisista malleista on suora viiva, ja tunnetusti suora on täysin määrätty, kun sen vakiot tunnetaan.

Regressiosuoran yhtälö

Tarkastellaan muuttujien X ja Y havaintoarvoista muodostettuja havaintopareja $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$. Pistejoukkoon sovitetun regressiosuoran yhtälön $y = b_0 + b_1 x$ vakiot eli parametrit saadaan pienimmän neliösumman menetelmän avulla seuraavasti:

$$b_1 = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i{}^2) - (\sum x_i)^2}$$

$$b_0 = \frac{\sum y_i - b_1(\sum x_i)}{n} = \overline{y} - b_1 \overline{x}$$

Parametrien b_0 ja b_1 laskemisessa käytettävien kaavojen todistaminen sivuutetaan. Vakioita b_0 ja b_1 sanotaan **regressiokertoimiksi**. Vakio b_0 ilmoittaa pisteen, jossa suora leikkaa y-akselin, ja sillä on selkeä tulkinta vain silloin, kun x voi saada myös arvon nolla. Vakio b_1 taas ilmaisee suoran kulmakertoimen, josta käy ilmi, kuinka paljon ja mihin suuntaan y:n arvo muuttuu, kun x kasvaa yhden yksikön verran.

Huomautus

Edellä olevia kaavoja käyttämällä saadaan suoran yhtälö ratkaistuksi, olivatpa x:n ja y:n arvot **mitkä tahansa**. Sen tähden on aina aluksi tehtävä havaintoaineistosta hajontakuvio ja katsottava siitä, onko lainkaan mielekästä kuvata pisteparvea suoralla viivalla. Vasta jos vastaus on myönteinen, kannattaa laskea regressiosuoran yhtälön parametrit.

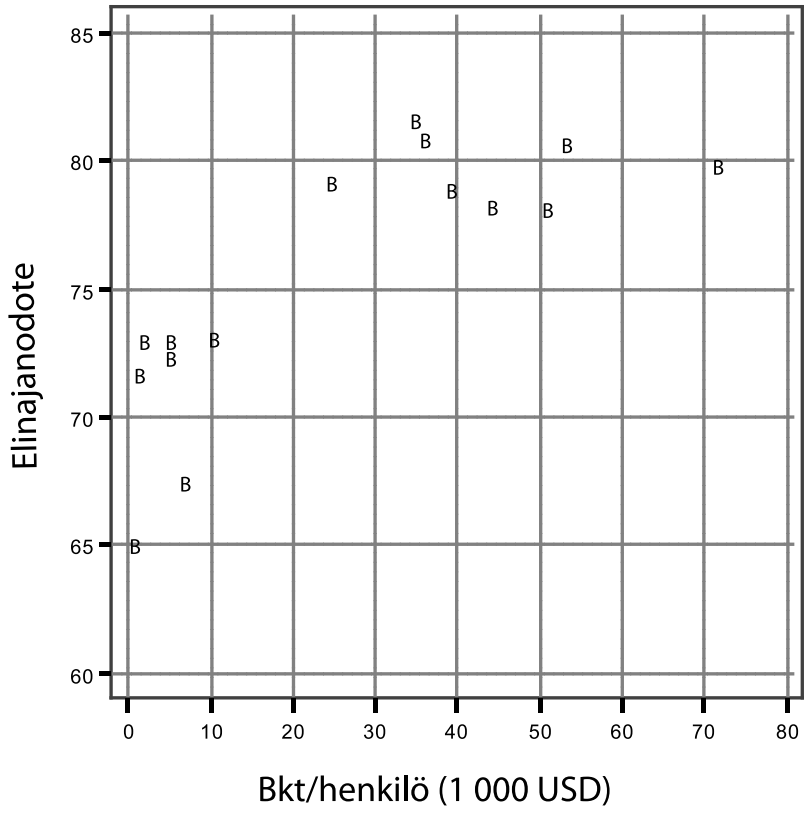
ESIMERKKI 1

Kun tutkittiin eri maiden eliniänodotteita, arveltiin, että ne ovat yhteydessä bruttokansantuotteeseen (bkt) siten, että bruttokansantuotteen kasvaessa myös elinajanodote kasvaa. Tämän olettamuksen selvittämiseksi kerättiin vuodelta 2006 tilasto, jossa esitetään bruttokansantuote henkeä kohti Yhdysvaltain dollareina sekä elinajanodote 0-vuotiaille (lähde: Tilastokeskus). Tulokset olivat seuraavat:

Maa	Bkt/henkilö (USD)	Elinajanodote
Australia	36 016	80,5
Brasilia	5 177	72,0
Egypti	1 432	71,3
Intia	769	64,7
Islanti	53 228	80,3
Japani	34 955	81,3
Kiina	1 944	72,6
Norja	71 674	79,5
Suomi	39 332	78,5
Tanska	50 808	77,8
Turkki	5 202	72,6
Unkari	10 298	72,7
Yhdysvallat	44 315	77,9
Uusi-Seelanti	24 566	78,8
Venäjä	6 861	67,1

Järkevyystarkastelun aluksi todetaan, että bruttokansantuotteella voi olla vaikutusta elinajanodotteeseen, sillä suuri bruttokansantuote mahdollistaa muun muassa hyvät terveyspalvelut. Toisaalta elinajanodotteeseen todennäköisesti vaikuttavat muutkin tekijät kuin pelkästään bruttokansantuote. Ei kuitenkaan ole mitään loogista estettä sille, ettei bruttokansantuotteella voisi olla vaikutusta myös elinajanodotteeseen.

Seuraavassa kuvataan havaintoaineisto hajontakuvion avulla.



Bruttokansantulo ja elinajanodote.

Yllä olevan kuvan perusteella voidaan yleisesti havaita, että elinajanodote kasvaa bruttokansantuotteen mukana. Silmämääräisesti suoraviivainen malli on siis mahdollinen kuvaamaan muuttujien välistä yhteyttä. Seuraavaksi lasketaankin suoran yhtälön parametrit. Tätä varten muutetaan bruttokansantuotteen yksiköksi 1 000 USD ja täydennetään tilastoa kahdella ylimääräisellä sarakkeella.

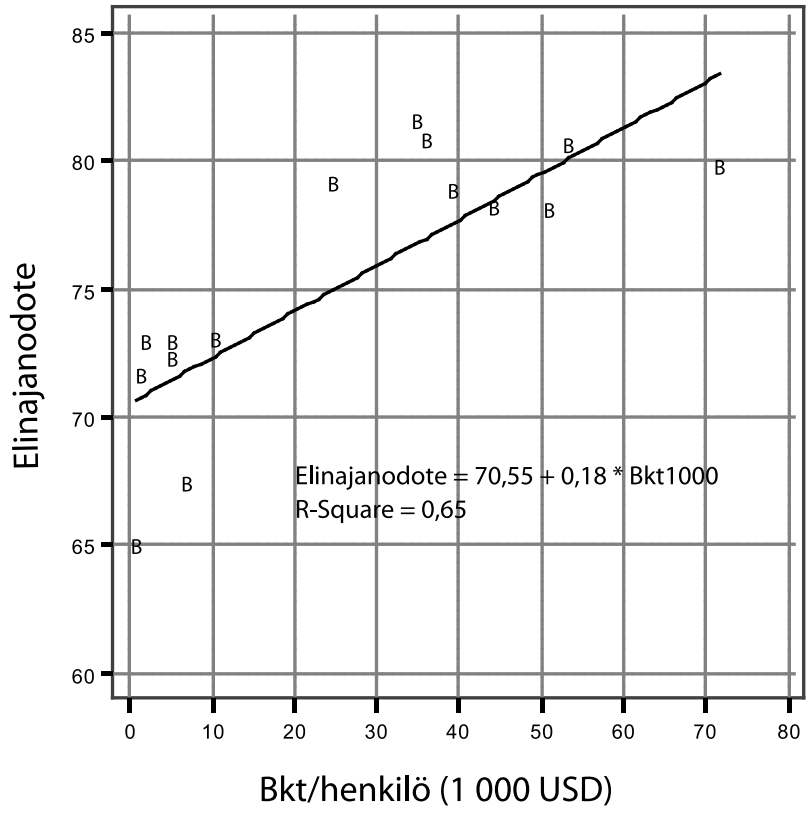
Maa	Bkt/henkilö (1 000 USD), x	Elinajan-odote, y	$x \cdot y$	x^2
Australia	36,02	80,5	2 899,288	1 297,152
Brasilia	5,18	72,0	372,744	26,801
Egypti	1,43	71,3	102,102	2,051
Intia	0,77	64,7	49,754	0,591
Islanti	53,23	80,3	4 274,208	2 833,220
Japani	34,96	81,3	2 841,842	1 221,852
Kiina	1,94	72,6	141,134	3,779
Norja	71,67	79,5	5 698,083	5 137,162
Suomi	39,33	78,5	3 087,562	1 547,006
Tanska	50,81	77,8	3 952,862	2 581,453
Turkki	5,20	72,6	377,665	27,061
Unkari	10,30	72,7	748,665	106,049
Yhdysvallat	44,32	77,9	3 452,139	1 963,819
Uusi-Seelanti	24,57	78,8	1 935,801	603,488
Venäjä	6,86	67,1	460,373	47,073
Yhteensä	386,58	1127,6	30 394,222	17 398,559

$$b_1 = \frac{15 \cdot 30\,394,222 - 386,58 \cdot 1127,6}{15 \cdot 17\,398,559 - 386,58^2} = 0,1794$$

$$b_0 = \frac{1127,6 - 0,1794 \cdot 386,58}{15} = 70,55$$

Suoran yhtälö on siis $y = 70,55 + 0,1794\,x$. Saatua mallia voidaan tulkita muun muassa siten, että kun bruttokansantuote henkilöä kohti kasvaa 1 000 USD, elinajanodote kasvaa 0,18 vuotta.

Piirretään nyt suora saatuun pistejoukkoon ja tarkastellaan kuvasta mallin toimivuutta. Itse pistejoukko kuvaa tilastollista yhteyttä muuttujien välillä, kun taas suoran avulla mallinnetaan tätä yhteyttä.



Suora tilastoaineistoa kuvaavana mallina.

Tietokonetulosteen tulkinta

Tavallisesti regressioanalyysi tehdään tietokoneella, joka laskee käytetyn mallin parametreille arvot. Samalla tulostuu joitakin muita arvoja. Tulosteet poikkeavat eri ohjelmia käytettäessä toisistaan, mutta seuraavassa taulukossa on tyypillinen tuloste esimerkistä 1. Regressiokertoimien (regression coefficients) arvot käyvät ilmi tulosteesta. Niiden perusteella todetaan edellisessä esimerkissä laskettu tulos oikeaksi.

Coefficients[a]						
		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	70,550	1,246		56,603	,000
	Bkt/henkilö (1000 USD)	,1794	,037	,806	4,902	,000

[a] Dependent Variable: Elinajanodote

Esimerkin 1 tietokonetuloste.

Taulukon tulosteesta käyvät ilmi parametrien arvot sekä regressiokertoimien keskivirheet (standard error) eli arvioidut keskihajonnat. Lisäksi molemmille parametreille on laskettu tulosteeseen ***t*-arvo**, joka saadaan siten, että parametrin (taulukossa näkyvää arvoa tarkempi) arvo jaetaan keskivirheellään. Näin saatu *t*-arvo noudattaa Studentin *t*-jakaumaa, joten sitä voidaan käyttää testimuuttujana. Tulosteissa näkyy myös *p*-arvo sarakkeessa Sig. Tulosteisiin palataan vielä myöhemmin.

Ennusteiden laatiminen

Tulevaisuutta pyritään usein ennakoimaan. Ennakoinnit perustuvat malleihin eli skenaarioihin. Matemaattisen mallin avulla pyritään kuvaamaan ja selittämään tarkasteltavaa ilmiötä. Tärkeä syy matemaattisen mallin käyttämiselle on se, että sen avulla voidaan helposti laatia **ennusteita**. Ennusteet saattavat kuitenkin poiketa toisistaan monesta syystä, kuten käytettyjen lähtötietojen, käytetyn näkemyksen eli vision ja käytetyn mallin mukaan. On tärkeätä ymmärtää, että samaakin tilastoaineistoa voidaan kuvata monella erilaisella mallilla.

Mallit ja niiden avulla laaditut ennusteet saattavat muuttaa ihmisten käytöstä tai toimintatapoja. Yksi esimerkki tästä on kasvihuoneilmiön huomioiminen. Myös ennusteet maapallon tulevasta liikakansoittumisesta ovat havahduttaneet ihmiset huomaamaan, mitä on väistämättä edessä, jollei mitään muutosta tapahdu.

Kun matemaattinen malli on saatu valmiiksi, ennusteiden laatiminen sen avulla on vaivatonta. Saatuun malliin x:n paikalle sijoitetaan tarkasteltava arvo, jolloin mallista saadaan lasketuksi tarkasteltavaa x:n arvoa vastaava y:n **estimaatti** (estimate) eli y:n arvon ennuste. Mallin avulla laskettua ennustetta merkitään $\hat{y}$:llä.

ESIMERKKI 2

Laaditaan ennuste elinajanodotteelle, kun bruttokansantuote on 30 000 USD.

Sijoitetaan esimerkissä 1 saatuun suoran yhtälöön x:n paikalle luku 30 (huom.! yksikkönä on 1 000 USD) ja lasketaan vastaava $\hat{y}$:n arvo:

$$\hat{y} = 70{,}55 + 0{,}1794 \cdot 30$$

$$\hat{y} = 76.$$

Varminta on heti tarkistaa vastaus myös kuvasta. Tässä tapauksessa kuvasta voidaan lukea laskelmien kanssa yhtäpitävä tulos.

Koska kyseessä on ennuste, tulosta ei kannata antaa liian tarkasti. Niinpä voidaan sanoa, että jos maan bruttokansantuote henkilöä kohti on 30 000 USD, sen vastasyntyneiden lasten keskimääräinen elinajanodote on noin 76 vuotta.

Tulkinta: Kun verrataan saatua tulosta esimerkin 1 alkuperäisiin arvoihin ja sivun 265 kuvaan, käy ilmi, että tulos on mahdollinen ja mielekäs. Tulos on kuitenkin vain piste-estimaatti, joten sen luotettavuudesta ei ole tietoa. Tähän palataan kohdassa 9.6.

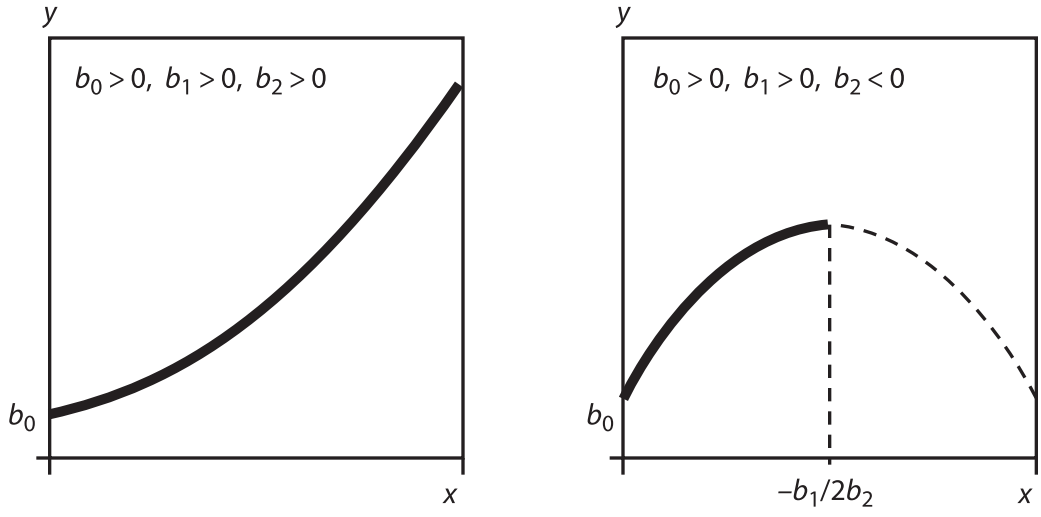
Ennusteen mekaaninen laatiminen on niin helppoa, että siitä tulee helposti rutiinitehtävä. Laatiminen on kuitenkin yleensä mahdollista vain silloin, kun ennusteeseen käytetty x:n arvo on melko lähellä muita x:n havaittuja arvoja. Ennusteet ovat aina ehdollisia. Niiden laadinta perustuu oletuksiin muun muassa siitä, että havaittu säännönmukaisuus on edelleen voimassa tai jatkuu tulevaisuudessa.

Ennustearvon laskemisen jälkeen on **aina** harkittava, onko saatu ennustearvo mahdollinen ja mielekäs. Ennusteen tarkkuutta käsitellään laajemmin kohdassa 9.5.

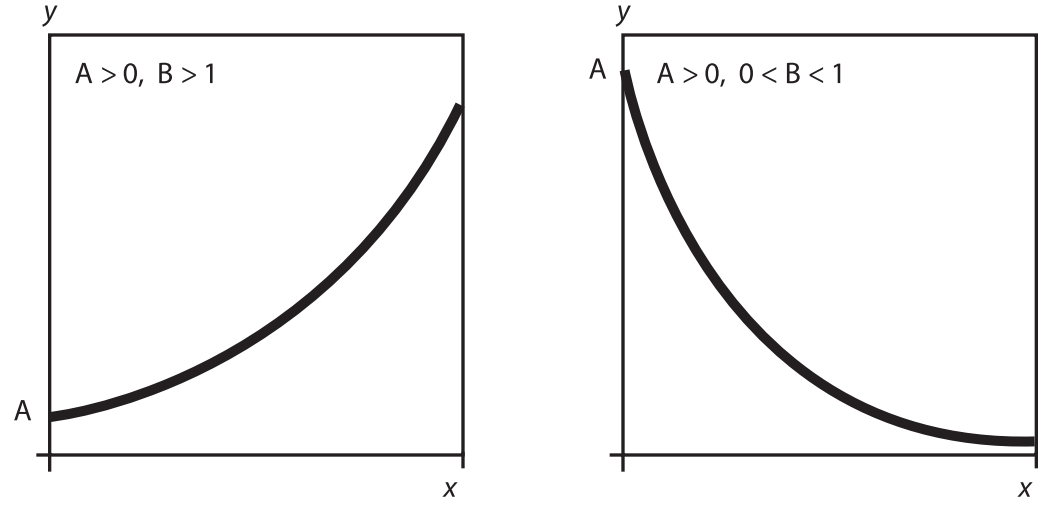
9.3 Muita malleja

Jos hajontakuvio osoittaa, että pistejoukko selvästi taipuu johonkin suuntaan, suoraviivaista mallia ei voida käyttää. Tällöin on harkittava esimerkiksi polynomimallia, eksponentiaalista mallia tai logaritmista mallia.

Malli: $y = b_0 + b_1x + b_2x^2$



Malli: $y = A \cdot B^x$



Matemaattisia malleja.

Joidenkin tilasto-ohjelmien avulla voidaan tietokone panna valitsemaan annettuun pistejoukkoon parhaiten sopiva malli. Useimmissa tilasto-ohjelmissa pitää ensin antaa mallin tyyppi, minkä jälkeen ohjelma tulostaa parametrien arvot.

Eksponentiaalinen malli

Eksponentiaalisen mallin tapauksessa mallin vakiot voidaan laskea ilman tilasto-ohjelmiakin pelkästään laskinta tai taulukkolaskentaohjelmaa käyttämällä. Eksponenttifunktio on muodoltaan

$Y = A \cdot B^x$.

Kun otetaan yhtälöstä logaritmit puolittain, saadaan

$\log Y = \log A + x \cdot \log B$.

Merkitsemällä $\log Y = y$, $\log A = b_0$ ja $\log B = b_1$ saadaan alkuperäinen eksponenttiyhtälö palautettua suoran yhtälöksi

$y = b_0 + b_1 x$.

Alkuperäisen yhtälön vakiot A ja B saadaan käänteislogaritmien avulla.

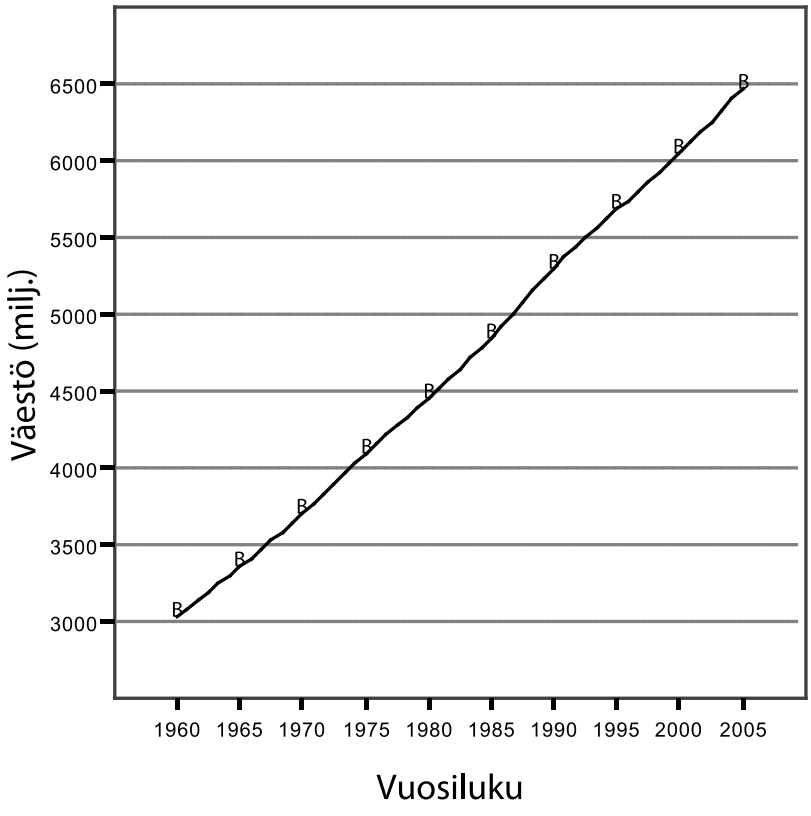
ESIMERKKI 3

Maapallon väkiluku (miljoonina) on kehittynyt seuraavasti:

Vuosi	Väestö (milj.)
1960	3 019
1965	3 347
1970	3 698
1975	4 084
1980	4 448
1985	4 844
1990	5 292
1995	5 674
2000	6 045
2005	6 465

Esitä kehitys eksponentiaalisen mallin avulla.

Piirretään tilastoaineistosta aluksi kuvaaja:



Maapallon väestön kehitys vuosina 1960–2005.

Maapallon väestöä kuvaava pistejoukko kaareutuu loivasti ylöspäin, joten eksponenttifunktio tuntuu mahdolliselta mallilta.

$$\text{Väestö} = A \cdot B^{\text{Vuosi}}$$

Ottamalla logaritmit (tässä on käytetty luonnollisia logaritmeja) puolittain saadaan

$$\ln(\text{Väestö}) = \ln(A) + \text{Vuosi} \cdot \ln(B).$$

Hieman pienemmillä luvuilla päästään, jos vuosiluvuista vähennetään 1900. Laskut näyttävät seuraavilta:

Vuosi	Väestö (milj.)	ln(Väestö)	Vuosi · ln(Väestö)	Vuosi^2
60	3 019	8,0126809	480,7608558	3 600
65	3 347	8,1158197	527,5282806	4 225
70	3 698	8,2155474	575,0883188	4 900
75	4 084	8,3148322	623,6124134	5 625
80	4 448	8,4002098	672,0167869	6 400
85	4 844	8,4854961	721,2671689	7 225
90	5 292	8,5739515	771,6556373	8 100
95	5 674	8,6436496	821,1467135	9 025
100	6 045	8,7069868	870,6986763	10 000
105	6 465	8,7741583	921,2866206	11 025
Yhteensä	46 916	84,243332	6985,061472	70 125

Kun lasketut summat sijoitetaan regressiokertoimien kaavaan, saadaan b_0 = ln(A)= 7,024869 ja b_1 = ln(B) = 0,0169632.

Laskemalla sitten käänteislogaritmit tai soveltamalla logaritmin määritelmää havaitaan, että

vakio A on exp(7,024869) = 1124,25

vakio B on exp(0,0169632) = 1,017108.

Sama tieto käy ilmi myös seuraavasta tietokoneohjelman tulostamasta väestötilaston taulukosta:

Dependent Variable: Väestö (milj.)

	Model Summary					Parameter Estimates	
Equation	R Square	F	df1	df2	Sig.	Constant	b1
Exponential	,994	1244,242	1	8	,000	1124,250	,0169632

The independent variable is Vuosiluku.

Näin malliksi saadaan

$$\text{Väestö} = 1\,124{,}25 \cdot 1{,}017108^{\text{Vuosi}}.$$

Tarkistuksen vuoksi lasketaan mallin antamat väestöennusteet joiltakin vuosilta (sulkeissa todelliset väestömäärät):

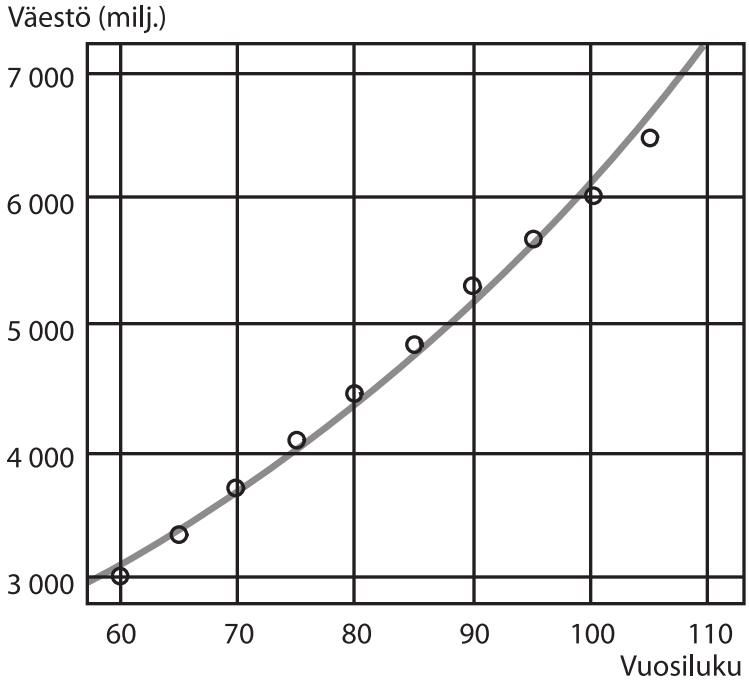
Vuosi 1960: $A \cdot B^{60} = 3\,111$ (3 019)

Vuosi 1970: $A \cdot B^{70} = 3\,686$ (3 698)

Vuosi 1980: $A \cdot B^{80} = 4\,367$ (4 448)

Vuosi 1990: $A \cdot B^{90} = 5\,175$ (5 292)

Vuosi 2000: $A \cdot B^{100} = 6\,132$ (6 045)



Eksponenttifunktio sovitettuna maapallon väestön kehitykseen vuosina 1960–2005.

Tulkinta: Vaikuttaa siltä, että malli kuvaa havaintoaineistoa melko hyvin. Kannattaa kuitenkin huomata, ettei eksponenttifunktiota voi käyttää mallina kovin pitkälle tulevaisuuteen, esimerkiksi vuoteen 2050. Väestön eksponentiaalisen kasvun raja tulee nimittäin vastaan viimeistään silloin, kun maapallolta ei saada enää tarpeeksi ravintoa kaikkien ihmisten ruokkimiseen.

Eksponenttifunktiota käytetään mallina monissa taloustieteellisissä sovelluksissa. Esimerkiksi jos pääoma, liikevaihto, tuotanto tms. kasvaa korkoa korolle tietyn kiinteän prosenttiluvun mukaan, tulosta kuvataan eksponenttifunktion avulla. Edellisessä esimerkissä oletetaan, että maapallon väestömäärä kasvaa vuosittain noin 1,7 %. Tulos näkyy vakion B arvosta. Todellisuudessa väestön kasvu on hieman tasaantunut, ja vuonna 2006 kasvuprosentti oli enää 1,17. Edellisessä mallissa kuitenkin oletetaan, että väestönkasvu pysyy ennallaan. Tämä käy ilmi myös kuviosta. Eksponenttimalli ei siis ehkä olekaan enää paras malli kuvaamaan maapallon väestön kehitystä. Mutta mikä voisi olla parempi malli?

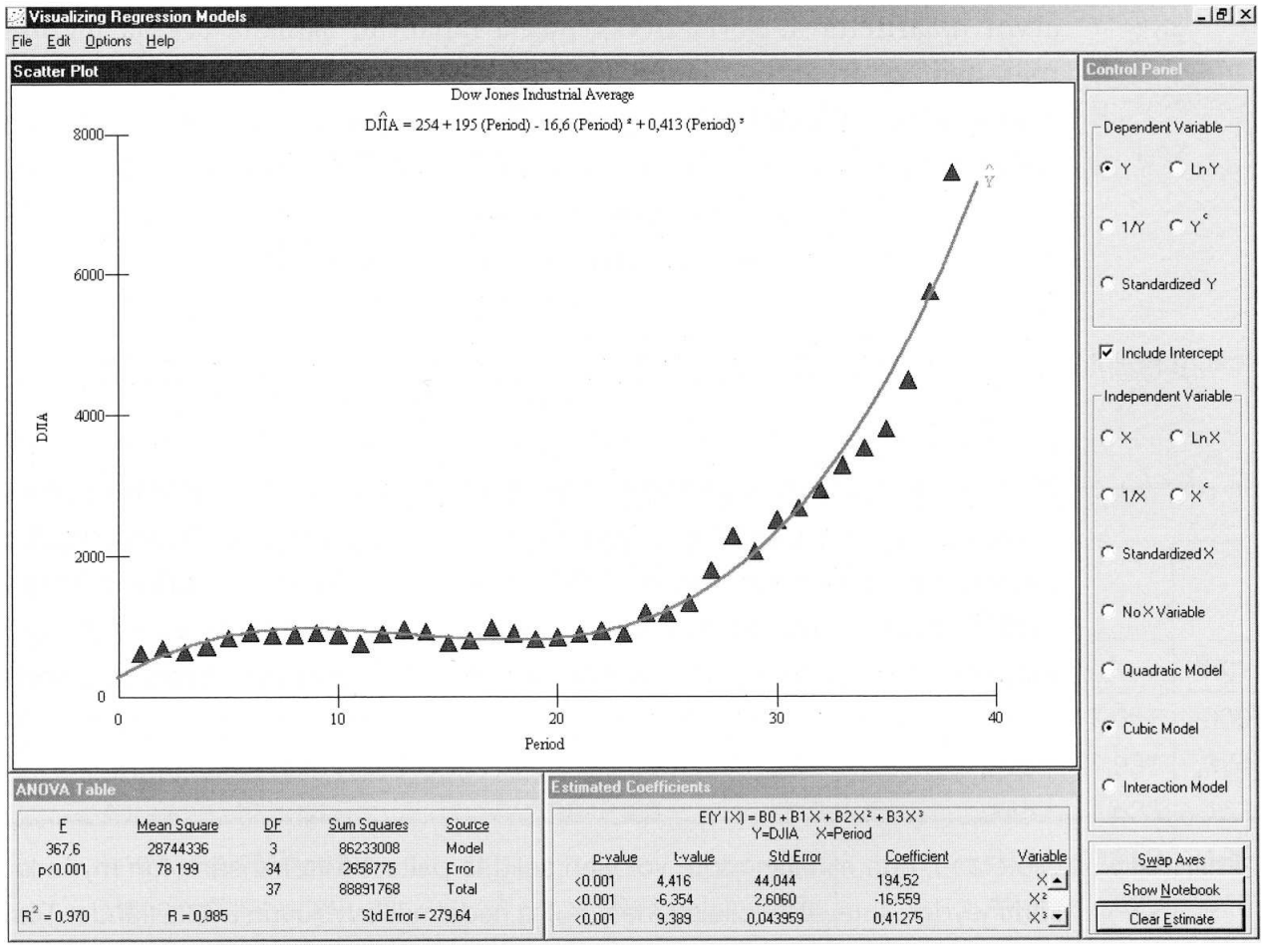
Polynomimalli

Polynomimalleista otetaan tavallisesti käyttöön toisen ja kolmannen asteen käyrät tai tarkemmin sanottuna jokin osa niistä. On mahdollista laskea näillekin malleille parametrien arvot laskimella tai taulukkolaskentaohjelmalla. Tavallisesti tehtävä kuitenkin tehdään jollakin tilastoohjelmalla. Laskentakaavoihin ei tämän vuoksi puututa tässä.

Jos hajontakuviosta voidaan hahmottaa maksimi, minimi ja käännepiste, pistejoukkoa kannattaa yrittää mallintaa esimerkiksi kolmannen tai viidennen asteen käyrällä.

ESIMERKKI 4

Pörssikurssien tutkiminen ja ennakoiminen on aina ollut vaikeaa. Seuraavassa kuvassa esitetään Dow Jones -indeksin peräkkäisiä arvoja. Pistejoukkoa on mallinnettu tietokoneohjelman avulla kolmannen asteen polynomilla. Kuvasta voidaan havaita, että mallintaminen on onnistunut hyvin.



Dow Jones -indeksin mallintaminen Visual Statistics -ohjelmalla.

9.4 Usean muuttujan mallit

Tähän mennessä on tarkasteltu tapauksia, joissa selitettävän muuttujan vaihteluita on selitetty vain yhden muuttujan avulla. Käytännössä selittäviä muuttujia on usein monia. Seuraavassa tarkastellaan, kuinka useiden selittävien muuttujien mukaan ottaminen tapahtuu.

Usean selittävän muuttujan lineaarinen regressiomalli on muodoltaan

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k.$$

Selittäviä muuttujia x_i valittaessa on pyrittävä siihen, etteivät ne korreloi keskenään. Jos jotkin selittävät muuttujat korreloivat selvästi keskenään, esiintyy **multikollineaarisuutta** (multicollinearity), jolloin muuttujat eivät juuri tuo lisää informaatiota ja on vaikea todeta, mikä on minkin muuttujan vaikutus selitettävään muuttujaan.

Mallin muodostamiseksi on määritettävä regressiokertoimien eli vakioiden b_0, b_1,... b_k arvot. Periaatteessa tämä tehdään samalla tavalla pienimmän neliösumman menetelmällä kuin yhden muuttujan tapauksessa. Sen sijaan laskennallisesti ero on suuri. Yhden muuttujan tapauksessa tavallisella laskimella voidaan laskut hoitaa kohtuullisella vaivalla, kun taas useamman muuttujan tapauksessa työmäärä on huomattavan suuri. Käytännössä laskelmat jätetäänkin tilasto-ohjelmien tehtäviksi.

Mallissa esiintyvät vakiot b_1, b_2,... b_k voidaan tulkita kunkin muuttujan vaikutuksen määrittäjiksi siten, että y muuttuu vakion b_i osoittamalla määrällä, kun muuttuja x_i lisääntyy yhden yksikön verran ja kaikkien muiden muuttujien arvot pysyvät vakioina. Regressiokertoimia ei kuitenkaan voi sellaisenaan tulkita selittävien muuttujien suhteellisen tärkeyden mitaksi. Kun selittävän muuttujan mittayksikkö muutetaan toiseksi (esimerkiksi eurot miljooniksi euroiksi), myös vastaava regressiokerroin muuttuu suuruudeltaan toiseksi.

ESIMERKKI 5

Oletetaan, että elinajanodote voidaan selittää paitsi bruttokansantuotteen myös lääkäritiheyden, imeväiskuolleisuuden (‰) ja hedelmällisyysluvun avulla (lähde: Tilastokeskus). Tilastot ovat vuodelta 2006, paitsi muuttujan henkilöä/lääkäri arvot, jotka ovat vaihtelevasti joitakin vuosia vanhempia.

Maa	Elinajan-odote	Bkt/henkilö (USD)	Henkilöä /lääkäri	Imeväis-kuolleisuus	Hedelmälli-syysluku
Australia	80,5	36 016	375	4,6	1,8
Brasilia	72,0	5 177	485	28,6	1,9
Egypti	71,3	1 432	448	31,3	2,8
Intia	64,7	769	1853	54,6	2,7
Islanti	80,3	53 228	360	3,3	1,9
Japani	81,3	34 955	489	3,2	1,4
Kiina	72,6	1 944	687	23,1	1,7
Norja	79,5	71 674	370	3,7	1,8
Suomi	78,5	39 332	313	3,6	1,7
Tanska	77,8	50 808	365	4,5	1,7
Turkki	72,6	5202	826	39,7	1,9
Unkari	72,7	10 298	260	8,4	1,3
Yhdysvallat	77,9	44 315	338	6,4	2,1
Uusi-Seelanti	78,8	24 566	306	5,8	1,8
Venäjä	67,1	6 861	208	15,1	1,3

Muodostetaan usean muuttujan regressiomalli.

Kun regressioanalyysi tehdään tietokoneella, saadaan seuraavan taulukon mukainen tuloste:

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	72,85802	3,872		18,817	,000
	Henkilöä/lääkäri	,00142	,004	,112	,359	,727
	Imeväiskuolleisuus	–,25736	,162	–,809	–1,589	,143
	Hedelmällisyysluku	2,14967	2,669	,180	,805	,439
	Bkt/henkilö (1000 USD)	,06401	,067	,287	,949	,365

[a] Dependent Variable: Elinajanodote

Useamman muuttujan regressioanalyysin tuloste 1.

Taulukosta käy ilmi, että usean muuttujan regressiomalli on

$y = 72{,}858 + 0{,}00142x_1 - 0{,}257x_2 + 2{,}150x_3 + 0{,}0640x_4$,

jossa ensimmäinen muuttuja x_1 on henkilöä/lääkäri ja vastaavasti viimeinen muuttuja x_4 on bruttokansantuote henkilöä kohti (tuhansia USD).

Tulkinta: Lääkäritiheys, hedelmällisyysluvut ja bruttokansantuote pidentävät elinajanodotetta, kun taas imeväiskuolleisuus lyhentää sitä. Saatua mallia tarkastellaan lähemmin kohdassa 9.5.

Huomautus

Lopullisessa mallissa voi olla vain niitä muuttujia, joita siihen on valittu. Tämä tarkoittaa sitä, että valitsemalla toisen joukon muuttujia voidaan saada aikaan toinen tulos ja mahdollisesti vielä parempi malli.

9.5 Mallin hyvyys

Kun muuttujien välistä yhteyttä kuvaava matemaattinen malli on saatu rakennetuksi ja mallin parametrien arvot lasketuiksi, on vielä selvitettävä, onko malli riittävän hyvä. Tällöin tutkitaan, onko syytä lisätä tai poistaa selittäviä muuttujia ja onko mallin muotoa muutettava vai onko saavutettu riittävän hyvä malli. Tämän selvittämiseksi on kehitetty useita menetelmiä.

Selityskerroin

Mallin hyvyys vaikuttaa osaltaan siihen, kuinka luotettavina sen avulla laskettuja ennusteita voidaan pitää. Yksi tapa arvioida ennusteen luotettavuutta on laskea **selityskerroin** eli **selitysaste** (coefficient of determination), joka mittaa mallin kykyä kuvata selitettävän muuttujan vaihtelua. Suuri selityskerroin ei kuitenkaan takaa tarkkoja ennusteita.

Yhden selittävän muuttujan tapauksessa selityskerroin prosentteina saadaan lasketuksi korrelaatiokertoimen r avulla seuraavasti:

selityskerroin $R^2 = r^2 \cdot 100\ \%$.

Selityskerroin ilmaisee, kuinka monta prosenttia selitettävän muuttujan y arvojen vaihteluista voidaan selittää selittävän muuttujan x avulla. Jos selityskerroin on pieni, pelkkä x ei selitä paljoakaan y:n arvojen vaihtelusta vaan valtaosa siitä johtuu muusta kuin x:stä. Jos taas selityskerroin on suuri, muuttuja x yksin selittää suurimman osan y:n arvojen vaihtelusta, jolloin malli kuvaa hyvin aineistoa.

Esimerkiksi elinajanodotetta kuvaavassa esimerkissä 1 korrelaatiokerroin $r = 0{,}806$, josta selityskertoimeksi saadaan 65 %. Tällöin siis 65 % elinajanodotteen vaihtelusta voidaan selittää bruttokansantuotteen perusteella.

Jos kyseessä on useamman muuttujan regressiomalli, selityskerroin R^2 ilmaisee, kuinka monta prosenttia muuttujan y arvojen vaihtelusta voidaan selittää yhteisesti kaikkien selittävien muuttujien x_i avulla. Pelkkä R on **yhteiskorrelaatiokerroin**, joka ilmaisee, kuinka voimakas yhteys vallitsee selitettävän muuttujan y ja kaikkien selittävien muuttujien x_i välillä.

Jotkin tilasto-ohjelmat laskevat tavallisen selityskertoimen lisäksi **suhteutetun selityskertoimen** (adjusted coefficient of determination) R_a^2 tai R^2(adj), joka ottaa lisäksi huomioon selittävien muuttujien lukumäärän. Tavallinen selityskerroin kasvaa selittävien muuttujien lukumäärän kasvaessa, mutta suhteutettu selityskerroin kasvaa vain, jos uusi selittävä muuttuja parantaa mallia. Tämän vuoksi on yleensä niin, että mitä suurempi R_a^2 on, sitä parempi malli on. Tätä seikkaa voidaan käyttää yhtenä kriteerinä, kun valitaan selittäviä muuttujia.

T-arvot

Mallin hyvyyttä tutkitaan usein t- tai p-arvojen avulla. Tietokonetulosteesta nähdään tavallisesti ainakin toinen näistä arvoista. ***t*-arvojen** avulla testataan hypoteesia, jonka mukaan kyseinen mallin parametri on nolla. Jos halutaan pikaisesti nähdä, poikkeaako parametrin arvo tilastollisesti nollasta, t-arvon tulisi olla vähintään 2. Tarkan arvon saa t-jakauman taulukosta, kun vapausasteita f (tai df) on $n–k$, jossa n = havaintoarvojen lukumäärä ja k laskettujen parametrien lukumäärä. Sivun 276 taulukosta käy ilmi, että vakiotekijä b_0 poikkeaa selvästi nollasta ($t = 18{,}817 > 2$) mutta kulmakerroin b_1 voi olla nolla ($t = 0{,}359 < 2$).

T-arvojen käyttöön tutustutaan esimerkin 6 avulla. Siinä tutkitaan useamman muuttujan regressiomallia, joka saatiin esimerkissä 5. Malliin valittiin useita selittäviä muuttujia, muttei tutkittu, mitkä muuttujat todella vaikuttavat tulokseen. Tässä käytettävä menetelmä on nimeltään **poistava valinta** (backward elimination).

Aluksi otetaan mukaan kaikki mahdolliset selittävät muuttujat, joilla uskotaan olevan vaikutusta tulokseen. Tämän jälkeen lasketaan tietokoneohjelman avulla mallin parametreille arvot ja niille *t*-arvot. Mallista poistetaan yksitellen aina se selittävä muuttuja, jonka *t*-arvo on itseisarvoltaan pienin. Yleensä vakiotekijää ei kuitenkaan poisteta, vaikka sen *t*-arvo olisikin itseisarvoltaan pienin.

Muuttujan poistamisen jälkeen lasketaan mallin parametreille uudet arvot ja niille *t*-arvot. Tämän jälkeen edellä mainittua menettelyä toistetaan, kunnes pieninkin *t*-arvo on itseisarvoltaan suurempi kuin ennalta asetettu raja, esimerkiksi 2. Näin päädytään malliin, jossa ovat mukana vain ne selittävät muuttujat, jotka todennäköisesti vaikuttavat selitettävään muuttujaan.

Se, mitä edellä sanottiin *t*-arvoista, pätee seuraavalla tavalla tulkittuna myös *p*-arvoihin: Suuri *t*-arvo vastaa pientä *p*-arvoa, ja pieni *t*-arvo vastaa suurta *p*-arvoa. Kun siis poistavassa valinnassa lähdettiin liikkeelle siitä selittävästä muuttujasta, jonka *t*-arvo oli itseisarvoltaan pienin, tilanne on samanlainen kuin silloin, kun ensin poistetaankin se selittävä muuttuja, jonka *p*-arvo on suurin. Kaikki muut vaiheet ovat täsmälleen samanlaisia. *t*:lle ennalta asetettua rajaa 2 vastaa suunnilleen *p*:n arvo 0,05 eli 5 %.

Jotkin tilasto-ohjelmat tulostavat selittävien muuttujien **VIF-arvot** (variance inflatory factor) (katso esim. kuvaa s. 287), joiden avulla pyritään poistamaan selittävien muuttujien joukosta ne muuttujat, joiden välillä on multikollineaarisuutta. Nyrkkisääntönä pidetään sitä, ettei VIF-arvo saisi olla yli 5. Jos useiden selittävien muuttujien VIF-arvo on yli 5, ensiksi poistetaan se muuttuja, jonka VIF-arvo on suurin. Näin jatketaan, kunnes jokaisen selittävän muuttujan VIF-arvo on alle 5.

ESIMERKKI 6

Tutkitaan poistavan valinnan menetelmällä, mitkä esimerkin 5 selittävistä muuttujista todennäköisesti vaikuttavat elinajanodotteeseen. Asian selvittämiseksi tarkastellaan aluksi sivun 276 taulukon tulostetta, josta käy ilmi, että itseisarvoltaan pienin *t*-arvo on lääkäritiheydellä. Koska muuttujan henkilöä/lääkäri *t*-arvo on 0,359 < 2, se poistetaan selittävien muuttujien joukosta ja tehdään uusi tuloste. Näin jatkamalla saadaan seuraava taulukko:

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	72,85802	3,872	,112	18,817	,000
	Henkilöä/lääkäri	,00142	,004	–,809	,359	,727
	Imeväiskuolleisuus	–,25736	,162	,180	–1,589	,143
	Hedelmällisyysluku	2,14967	2,669	,287	,805	,439
	Bkt/henkilö (1000 USD)	,06401	,067		,949	,365
2	(Constant)	72,79994	3,712		19,611	,000
	Imeväiskuolleisuus	–,21256	,099	–,668	–2,145	,055
	Hedelmällisyysluku	2,01973	2,538	,169	,796	,443
	Bkt/henkilö (1000 USD)	,07655	,055	,344	1,383	,194
3	(Constant)	75,11426	2,272		33,066	,000
	Imeväiskuolleisuus	–,15629	,068	–,491	–2,286	,041
	Bkt/henkilö (1000 USD)	,09766	,048	,439	2,041	,064

[a] Dependent Variable: Elinajanodote

Useamman muuttujan regressioanalyysin tuloste 2.

Taulukon tulosteesta käy ilmi, että jäljelle jääneillä selittävillä muuttujilla *t*-arvojen itseisarvo on vähän yli kaksi, joten muuttujien poistaminen lopetetaan. Näin päädytään malliin

$$y = 75{,}114 - 0{,}1563x_1 + 0{,}09766x_2,$$

jossa y on elinajanodote, x_1 imeväiskuolleisuus (‰) ja x_2 bruttokansantuote henkilöä kohden (1 000 USD).

Tulkinta: Mallin mukaan bruttokansantuotteen kasvu lisää elinajanodotetta ja imeväiskuolleisuuden kasvu vähentää elinajanodotetta. Molemmat johtopäätökset tuntuvat erikseen järkeviltä. On kuitenkin huomattava, että tässä malliin valittiin mukaan vain joitakin mahdollisia selittäviä muuttujia. Jos mukaan olisi otettu muita selittäviä muuttujia, tulos olisi saattanut olla toinen. Selittävien muuttujien valinnalla on siis keskeinen merkitys.

Tutkitaan vielä, olisiko edellä saatu malli pystytty ennakoimaan korrelaatiokertoimien avulla. Seuraavaan taulukkoon on laskettu selitettävän muuttujan (= elinajanodote) ja kaikkien alkuperäisten selittävien muuttujien välinen korrelaatiokerroin:

Coefficients[a]

		Elinajan-odote	Hen-kilöä/ lääkäri	Imeväis-kuolleisuus	Hedel-mälli-syysluku	Bkt/hen-kilö (1000 USD)
Elinajan-odote	Pearson Correlation	1	–,566*	–,819**	–,326	,806**
	Sig. (2-tailed)		,028	,000	,235	,000
	N	15	15	15	15	15
Henkilöä/ lääkäri	Pearson Correlation	–,566*	1	,820**	,580*	–,415
	Sig. (2-tailed)	,028		,000	,023	,124
	N	15	15	15	15	15
Imeväis-kuolleisuus	Pearson Correlation	–,819**	,820**	1	,628*	–,747**
	Sig. (2-tailed)	,000	,000		,012	,001
	N	15	15	15	15	15
Hedelmälli-syysluku	Pearson Correlation	–,326	,580*	,628*	1	–,221
	Sig. (2-tailed)	,235	,023	,012		,428
	N	15	15	15	15	15
Bkt/henkilö (1000 USD)	Pearson Correlation	,806**	–,415	–,747*	–,221	1
	Sig. (2-tailed)	,000	,124	,001	,428	
	N	15	15	15	15	15

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

Korrelaatiomatriisi.

Taulukon ensimmäiseltä riviltä voidaan havaita, ettei yksikään korrelaatiokerroin ole itseisarvoltaan lähellä nollaa ja hedelmällisyyslukua lukuun ottamatta korrelaatiokerroin poikkeaa tilastollisesti nollasta. Tästä päätellään, ettei pelkän korrelaatiomatriisin perusteella voida selvittää, mitkä selittävät muuttujat vaikuttavat todennäköisesti lopulliseen malliin ja siten siihen, mitkä muuttujat malliin pitää valita.

Lopullisen mallin löytämiseksi on muitakin vaihtoehtoja kuin poistava valinta. Voidaan menetellä myös siten, että malliin lisätään yksi selittävä muuttuja kerrallaan ja tutkitaan sen jälkeen *t*-arvoja. Jos uuden muuttujan *t*-arvo on riittävän suuri, kyseinen muuttuja lisätään malliin. Muussa tapauksessa siirrytään seuraavaan mahdolliseen selittävään muuttujaan. Lähtökohdaksi valitaan se selittävä muuttuja, joka yksinään selittää parhaiten selitettävän muuttujan vaihtelua. Tätä menetelmää sanotaan **eteneväksi valinnaksi** (forward selection). Menetelmä ei kuitenkaan välttämättä anna samaa tulosta kuin poistava valinta.

Kolmantena vaihtoehtona on **askeltava regressioanalyysi** (stepwise regression analysis), jossa yhdistyvät molemmat edellä käsitellyt menetelmät. Poistavan valinnan menetelmässä kerran pois pudotettu muuttuja ei voi enää palata malliin. Etenevässä valinnassa taas kerran mukaan otettu muuttuja ei voi enää pudota pois. Askeltavassa regressioanalyysissä lähdetään liikkeelle samalla tavoin kuin etenevässä valinnassa eli ottamalla ensimmäiseksi selittäväksi muuttujaksi paras yhden selittäjän muuttuja. Tähän selittäjään lisätään seuraava selittäjä mutta jokaisen vaiheen jälkeen tutkitaan uudelleen, tulisiko kerran malliin valittu muuttuja kuitenkin pudottaa sieltä pois. Näin jatketaan, kunnes kaikki mahdolliset selittävät muuttujat on käyty läpi.

Jäännöstermin merkitys

Tarkastellaan vielä esimerkin 1 tilannetta. Jos yhden muuttujan lineaarista mallia käyttämällä lasketaan bruttokansantuotteen jokaisella x_i:n arvolla arvioitu elinajanodote $\hat{y}_i$, huomataan todellisten ja arvioitujen lukujen välillä eroja. Kun lukujen erotusta merkitään e:llä, eli $y - \hat{y} = e$, alkuperäinen malli muuttuu muotoon

$$y = b_0 + b_1x + e.$$

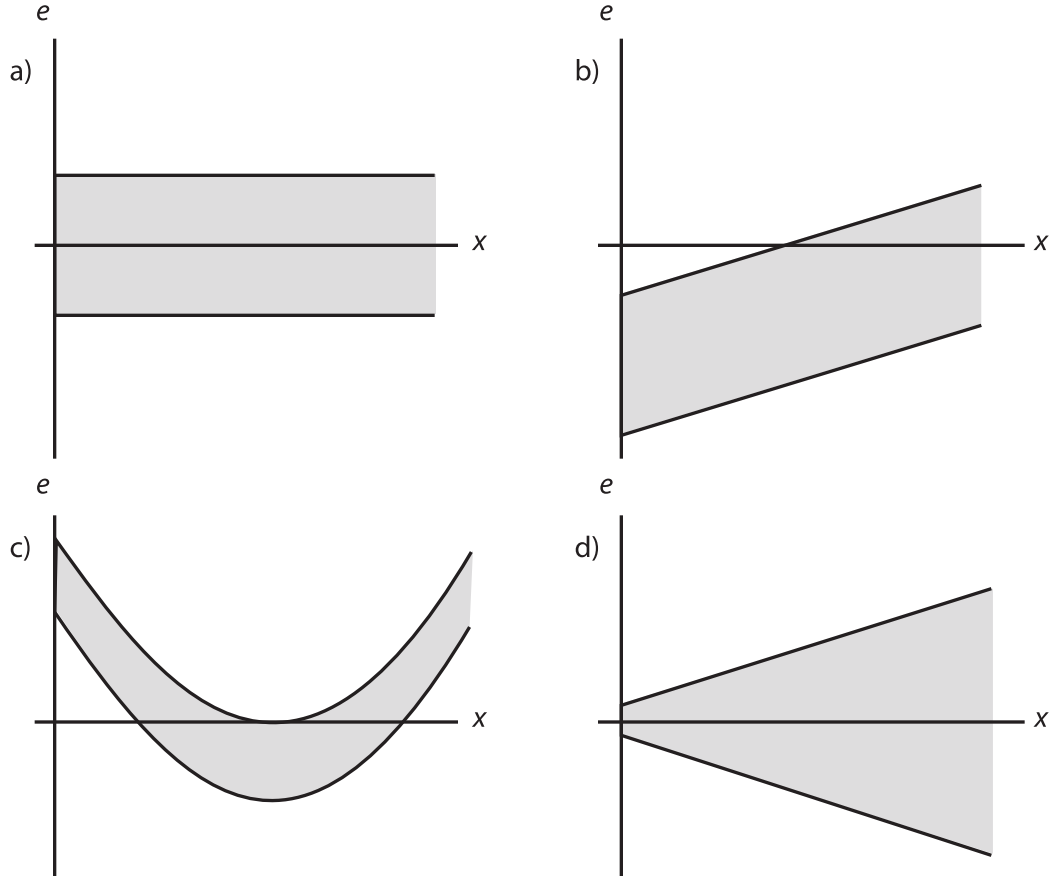
Elinajanodotetta (y) pyritään siis selittämään suoraviivaisesti bruttokansantuotteen (x) avulla lausekkeella $\hat{y} = b_0 + b_1x$. Lauseke $b_0 + b_1x$ kuvastaa **mallin determinististä osaa**. Malliin sisältyy kuitenkin satunnainen epävarmuustekijä (e), mallin **jäännöstermi** (residual), joka kuvastaa **mallin avulla selittämättä jäänyttä vaihtelua**. Jäännöstermi sisältää muun muassa malliin kuulumattomien muuttujien vaikutuksen, mittaus- tai tallennusvirheet ja satunnaisvaihtelun, ja sen avulla voidaan selvittää, kuinka hyvin tai huonosti valittu malli kuvaa havaintoaineistoa.

Diagnostiikka

Regressioanalyysissä diagnostiikalla pyritään selvittämään muun muassa mallin mahdollista väärää määrittelyä ja poikkeavien arvojen olemassaoloa.

Matemaattisen mallin sopivuutta tutkitaan usein jäännöstermin avulla. Tätä varten piirretään hajontakuvio siten, että pystyakselilla muuttujana on jäännöstermi *e* ja vaaka-akselilla muuttujana on *x*. Ellei hajontakuviossa ole mitään selkeätä säännönmukaisuutta, malli on todennäköisesti käyttökelpoinen. Jos taas hajontakuviossa esiintyy säännönmukaisuutta, malli ei ole riittävän hyvä. Tällöin täytyy tutkia esimerkiksi muuntyyppisten mallien tai useamman muuttujan mallin mahdollisuutta.

Alla olevassa kuvassa esitetään joitakin tyypillisiä jäännöstermiin liittyviä hajontakuvioita. Kuviossa a) esitetään ideaalinen tilanne: pistejoukko sijaitsee tasaisesti nollan molemmin puolin, mutta se on muuten satunnainen eikä hajontakuviossa esiinny trendiä. Kuviot b) ja c) taas osoittavat, että malli on määritelty väärin, ja kuvio d) liittyy tilanteisiin, joissa selittävänä muuttujana on esimerkiksi aika ja jäännöstermin hajonta kasvaa aikaa myöten.

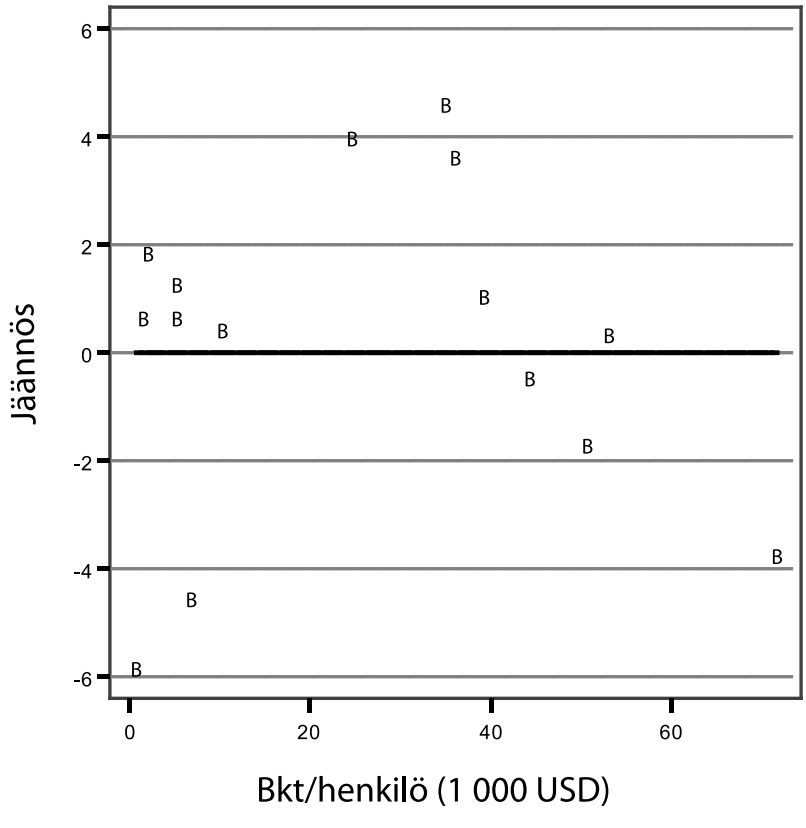


Jäännöstermeistä muodostettuja hajontakuvioita.

ESIMERKKI 7

Lasketaan esimerkin 1 jäännöstermin arvot eli jäännökset ja piirretään hajontakuvio.

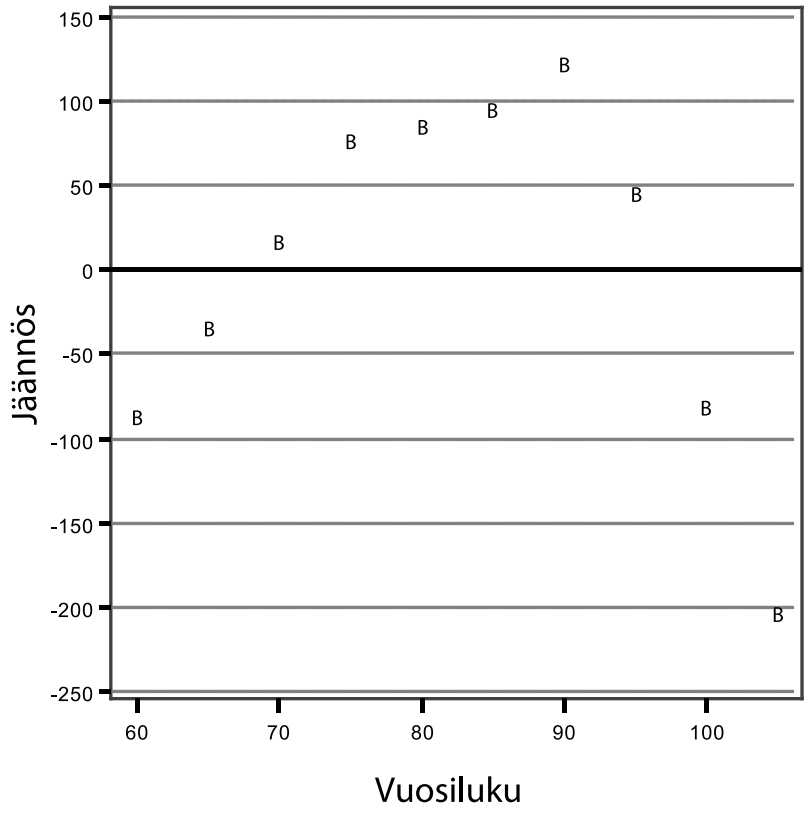
Maa	Bkt/henkilö (USD), x	Elinajanodote, y	$\hat{y}$	$e = y - \hat{y}$
Australia	36 016	80,5	77,0	3,5
Brasilia	5 177	72,0	71,5	0,5
Egypti	1 432	71,3	70,8	0,5
Intia	769	64,7	70,7	–6,0
Islanti	53 228	80,3	80,1	0,2
Japani	34 955	81,3	76,8	4,5
Kiina	1 944	72,6	70,9	1,7
Norja	71 674	79,5	83,4	–3,9
Suomi	39 332	78,5	77,6	0,9
Tanska	50 808	77,8	79,7	–1,9
Turkki	5 202	72,6	71,5	1,1
Unkari	10 298	72,7	72,4	0,3
Uusi-Seelanti	24 566	78,8	75,0	3,8
Venäjä	6 861	67,1	71,8	–4,7
Yhdysvallat	44 315	77,9	78,5	–0,6



Esimerkin 1 jäännöstermin hajontakuvio.

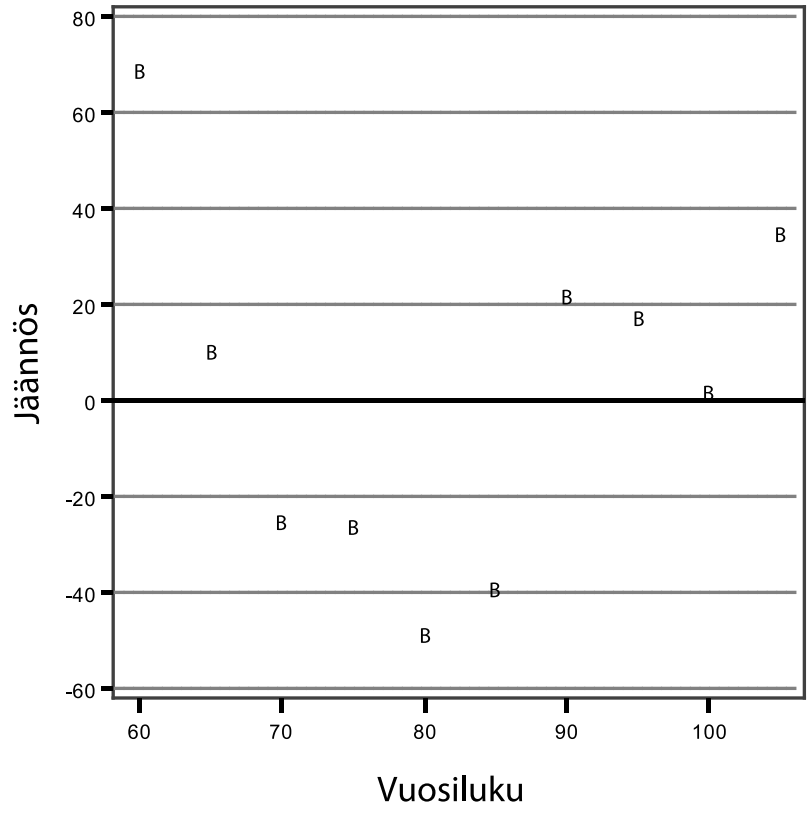
Tulkinta: Jäännöstermin hajontakuviosta voidaan havaita, että se muistuttaa jonkin verran edellisen sivun kuvan a-kohdan ideaalitapausta. Suoraviivainen malli näyttää siis hajontakuvion perusteella mahdolliselta. Tämä ei tietysti takaa sitä, että paras mahdollinen malli olisi löydetty. Ilmeisesti parempi malli löydettiin esimerkissä 6.

Maapallon väkiluvun kehitystä on vanhastaan kuvattu eksponenttifunktiolla. On mielenkiintoista todeta, että kun maapallon väkiluvun kehitystä mallinnetaan eksponenttifunktiolla, selityskerroin on suuri (99,4 %) mutta jäännöstermin hajontakuvio ei ole kovin hyvä (ks. alla olevaa kuvaa). Kuviossa näkyy selvää säännönmukaisuutta. Syynä on se, ettei väkiluvun lisäysprosentti ole pysynyt vakiona vaan se on pienentynyt erityisesti 1990-luvulta alkaen.



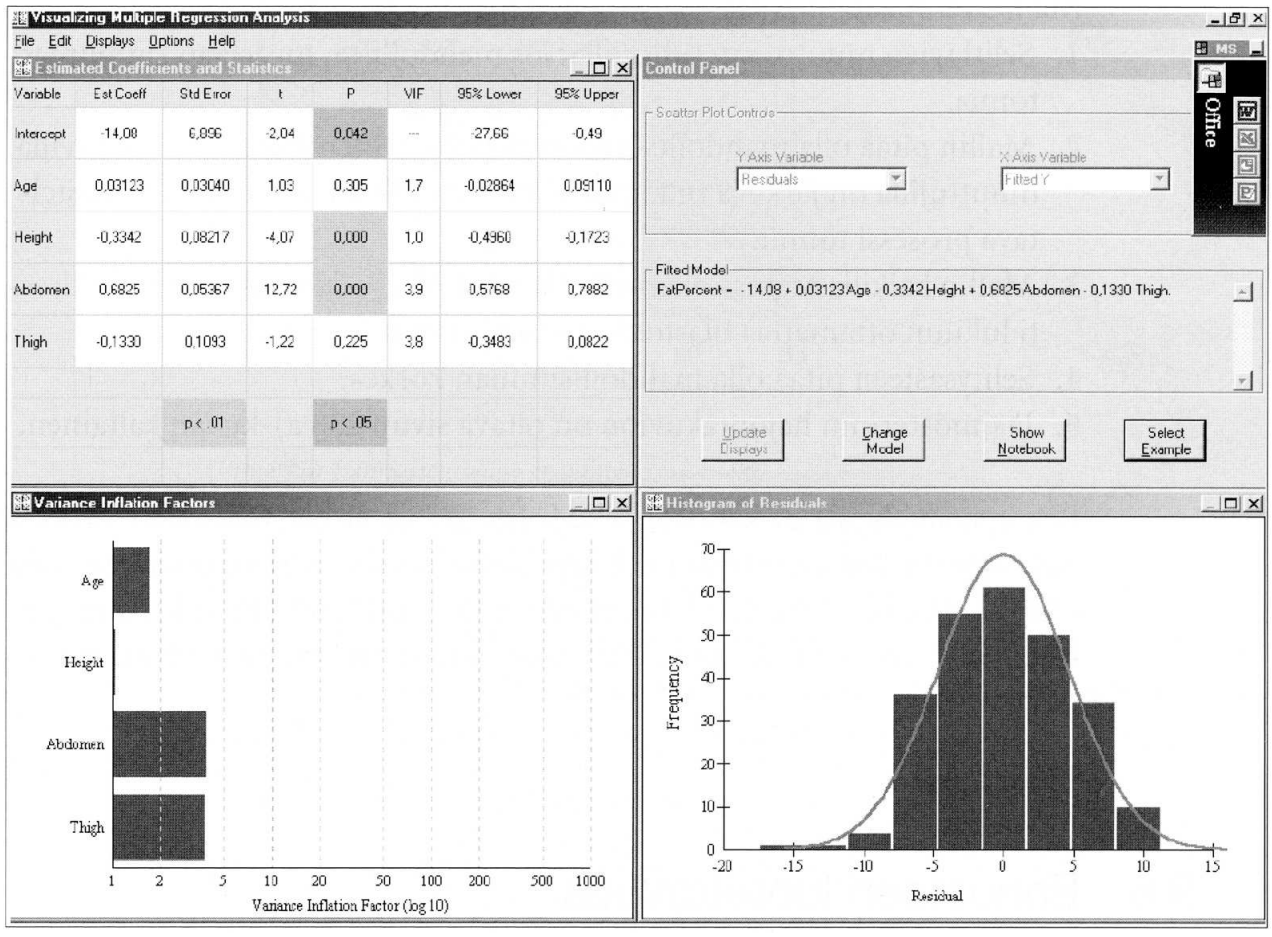
Jäännöstermin hajontakuvio, kun mallina on eksponenttifunktio.

Kun maapallon väkiluvun kehitystä mallinnetaan suoralla viivalla, jäännöstermin hajontakuvio (ks. seuraavaa kuvaa) on parempi kuin eksponenttifunktion hajontakuvio. Suora viivakin kelpaa siten malliksi, ja selityskerroin on jopa entistä suurempi (99,9 %).



Jäännöstermin hajontakuvio, kun mallina on suora.

Toinen vaihtoehto tutkia jäännöstermiä on luokitella sen arvot. Oletusten mukaan jäännöstermi jakautuu regressioanalyysissä normaalisti, joten luokituksen perusteella tulisi saada histogrammi, joka on yksihuippuinen ja melko symmetrinen siten, että se on keskittynyt nollan ympärille. Ellei näin ole, malli ei kuvaa havaintoaineistoa riittävän hyvin. Tämä menettely soveltuu käytettäväksi kuitenkin vain silloin, kun havaintoarvoja on melko runsaasti.



Mallin ja jäännöstermin tarkastelua Visual Statistics -ohjelmalla.

Poikkeavaksi arvoksi (outlier) sanotaan havaintoarvoa, joka poikkeaa huomattavasti muista vastaavista havaintoarvoista tai joka ei noudata muiden pisteiden edustamaa mallia. Poikkeava arvo saattaa syntyä esimerkiksi mittaus-, kirjoitus- tai kopiointivirheestä. Se voi kuitenkin olla aivan aito, oikea arvo, vaikkakin hyvin harvinainen. Poikkeava arvo erottuu esimerkiksi jäännöstermin hajontakuviossa äärimmäisenä arvona. Useat tilasto-ohjelmat varoittavat poikkeavista arvoista.

Jokainen poikkeava arvo tulisi tutkia erikseen. Jos arvo osoittautuu virheelliseksi, se on mahdollisuuksien mukaan korjattava oikeaksi tai poistettava aineistosta. Yhdelläkin poikkeavalla arvolla saattaa olla huomattava vaikutus tilastolliseen tutkimukseen.

Mallin valintaan liittyviä seikkoja

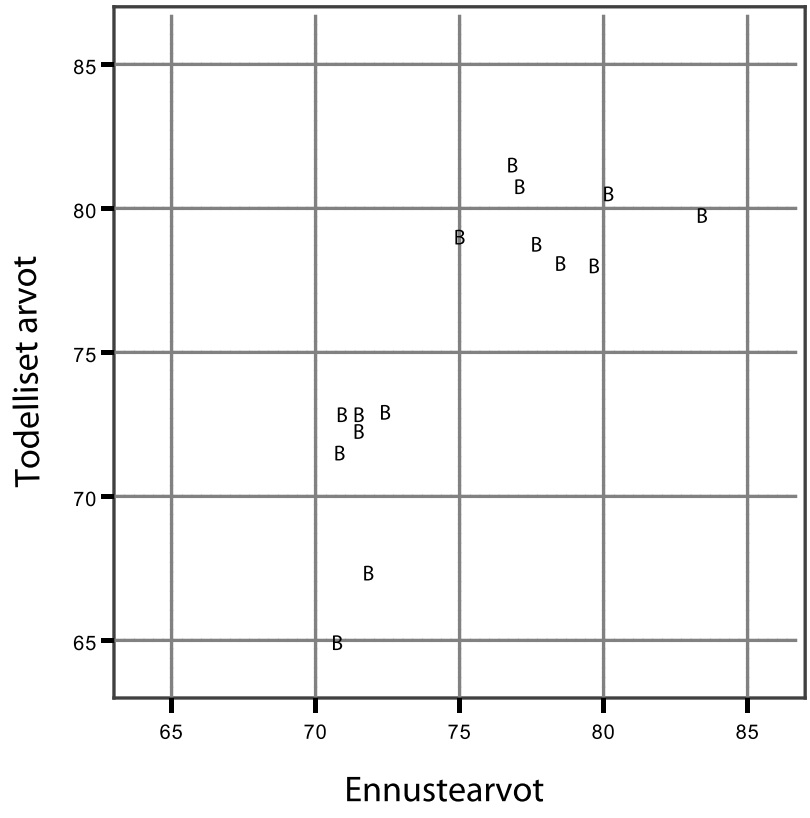
1. Selittävien muuttujien tulisi mahdollisuuksien mukaan olla toisistaan riippumattomia.
2. Mallin pitää olla mahdollisimman yksinkertainen siten, että selittävillä muuttujilla on järkevä tulkinta. Lisäksi malli kuvaa, kuinka tarkasteltava prosessi toimii.
3. Mallin kaikilla regressiokertoimilla pitää olla (vakiotekijää mahdollisesti lukuun ottamatta) tilastollisesti merkitsevät *t*-arvot.
4. Selityskertoimen pitää olla mahdollisimman suuri.
5. Jäännöstermin hajontakuvion on oltava sivun 283 a)-kuvan kaltainen.

Regressioanalyysiä ei kannata tehdä, ellei tietoja ole saatavilla vähintään viidestä havainnosta. Jos muuttujia on k kappaletta, havaintojen lukumäärän n on oltava suurempi kuin k. Mitä suuremmaksi erotus $(n–k)$ tulee, sitä tarkemmiksi parametrien estimaatit tulevat.

9.6 Ennusteen luotettavuus

Graafinen tarkastelu

Ennusteen tarkkuutta voidaan arvioida silmämääräisesti piirtämällä kuvaaja, jossa vaaka-akselille tulevat ennustearvot $\hat{y}$ ja pystyakselille alkuperäiset y:n arvot. Jos malli kuvaa hyvin pistejoukkoa, sen avulla lasketut arvot ovat lähellä y:n todellisia arvoja eli $\hat{y} \approx y$. Malli ja todellisuus vastaavat täysin toisiaan, jos $\hat{y} = y$ eli kaikki pisteet $(\hat{y}_i , y_i)$ sijaitsevat samalla suoralla. Käytännössä tätä ideaalitapausta ei kuitenkaan saavuteta. Edellä olevan perusteella voidaan todeta, että ennusteisiin sisältyy sitä enemmän epävarmuutta, mitä kauemmaksi pisteet $(\hat{y}_i , y_i)$ joutuvat ideaalitapausta esittävästä suorasta. Esimerkissä 1 lasketulle mallille saadaan seuraavassa kuvassa esitettävä graafinen tarkastelu.



Esimerkin 1 ennustearvojen tarkastelu.

Kuvasta käy ilmi, etteivät pisteet sijaitse kovinkaan lähellä kuvioon piirrettävää lävistäjää. Tällöin myöskään ennustearvoista ei tule kovin tarkkoja vaan niihin sisältyy epävarmuustekijöitä erityisesti pistejoukon alku- ja loppupäässä.

Ennusteväli ja ennustevyöhyke

Kun pienimmän neliösumman menetelmässä lasketaan regressiokertoimille b_0, b_1,... b_k arvot, muuttujien jakaumille ei aseteta mitään vaatimuksia. Kun tarkoituksena on testata hypoteeseja tai laskea ennusteille luottamusvälejä, jäännöstermin arvojen on oltava toisistaan riippumattomia ja noudatettava ainakin likimain normaalijakaumaa.

Kun regressiomalli on saatu valmiiksi, sen avulla laaditaan tavallisesti ennuste. Ennuste laaditaan aikaisemman esityksen mukaisesti siten, että tarkastelun kohteena olevan x:n arvo x^* sijoitetaan malliin:

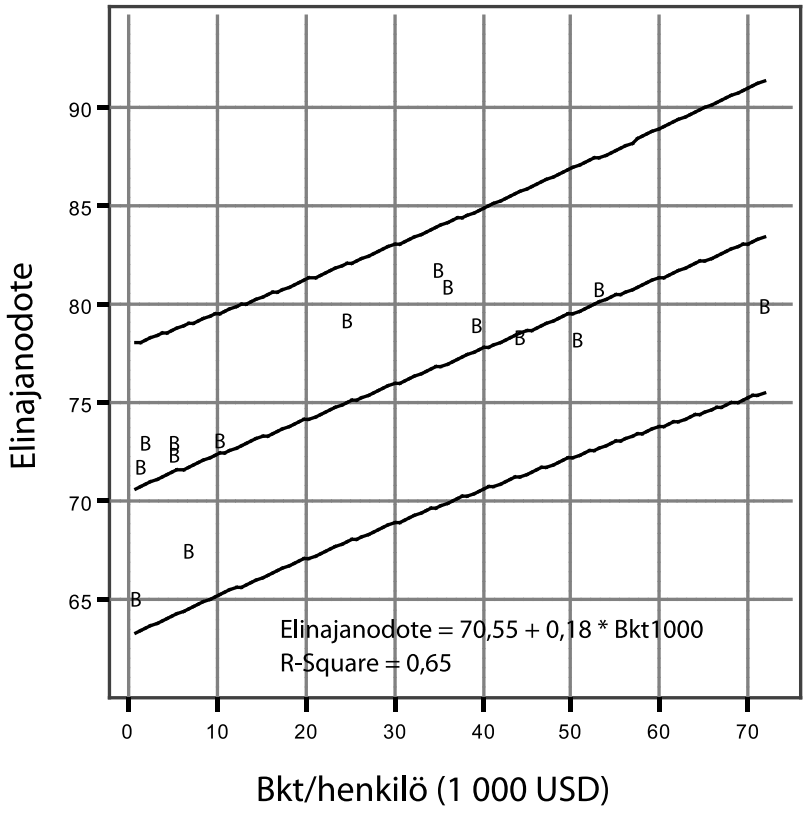
$$\hat{y}^* = b_0 + b_1 x^*$$

Mallin avulla lasketun ennusteen luotettavuutta selvitetään **ennustevälin** (prediction interval) avulla. Ennusteväli voidaan laskea jokaisella x:n arvolla. Tavallisesti tehtävä tehdään jollakin tilasto-ohjelmalla, minkä vuoksi laskentakaavoihin ei tässä puututa. Jos ennustevälin alarajat ja vastaavasti ylärajat yhdistetään käyrällä, regressiosuoran ympärille saadaan muodostetuksi **ennustevyöhyke** (prediction band).

Oheisessa kuvassa esitetään esimerkissä 1 lasketun mallin ennusteelle 95 %:n ennustevyöhyke. Kuvasta voidaan huomata, että ennustevyöhyke levenee hieman kuvion molemmista päistä. Ennustevyöhyke on kapeimmillaan silloin, kun $x = \bar{x}$.

Huomautus

Ennusteväli laaditaan muuttujille, kun taas luottamusväli muodostetaan parametreille (vakioille).



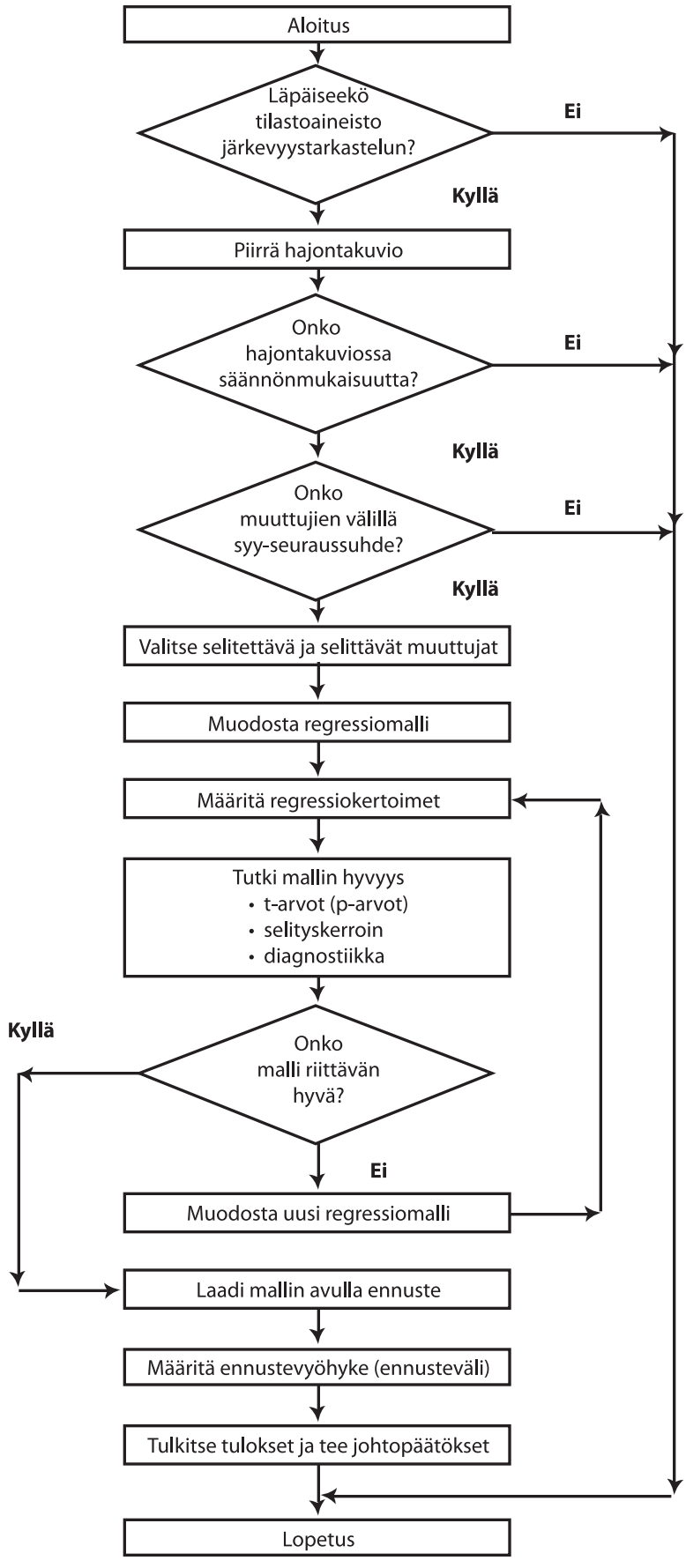
Esimerkin 1 ennustevyöhyke.

Ennustevyöhyke saadaan useimpien tilasto-ohjelmien avulla määrätyksi automaattisesti. Kuviosta ennusteväli saadaan määrätyksi likimääräisesti, mutta käytännössä riittävän tarkasti. Huom.! Tässä kaikki hajontakuvion pisteet ovat ennustevyöhykkeen sisällä.

Ennustevyöhykekuviosta voidaan havaita esimerkiksi seuraavaa. Jos maan bruttokansantuote on 30 000 USD, vastasyntyneen lapsen todennäköisin elinajan odote on 76 vuotta mutta 95 %:n todennäköisyydellä se on 68 ja 83 vuoden välillä. Ennusteeseen sisältyy aina väkisin paljon epävarmuutta, joka sitten heijastuu ennustevälin laajuutena. Erityisesti on huomattava, että vaikka mallin selityskerroin oli 65 %, se ei vielä taannut tämän kapeampaa ennusteväliä.

9.7 Regressiomallin vuokaavio

Regressiomallin rakentaminen voidaan esittää yksinkertaistettuna vuokaaviona seuraavalla sivulla esitetyn mukaisesti.



Regressiomallin vuokaavio.

9.8 Case C: Polttoaineen kulutus

Tutkimuksen tausta

Autoilijoita kiinnostaa autojen polttoaineen kulutus, koska se vaikuttaa välittömästi niiden käyttökustannuksiin. Arveltiin, että polttoaineen kulutukseen voisivat vaikuttaa moottorin teho, moottorin koko eli sylinteritilavuus, auton oma massa, kiihtyvyys, vaihteiden lukumäärä ja ilmanvastuskerroin. Luultavasti kaikkien tekijöiden vaikutusta ei voida helposti todentaa, mutta etukäteen ajateltuna ainakin jotkin edellä luetelluista tekijöistä voivat selittää polttoaineen kulutusta.

Tutkimus päätettiin rajata bensiinikäyttöisiin henkilöautoihin. Tilastoaineistona käytettiin uusien autojen esitteissä ja aikakauslehtien testeissä ilmoitettuja arvoja. Mukaan valittiin 13 yleistä automerkkiä: Audi, Citroën, Fiat, Ford, Honda, Hyundai, Kia, Lexus, Opel, Peugeot, Toyota, Volkswagen ja Volvo. Näistä joihinkin saatiin mukaan useita eri malleja. Polttoaineen kulutus ilmoitettiin yhdistettynä kulutuksena, jossa olivat mukana sekä kaupunkiajo että maantieajo.

Ilmanvastuskerroin ja vaihteiden lukumäärä jouduttiin lopulta jättämään pois selittävien muuttujien joukosta, koska kaikista autoista ei saatu näitä tietoja. Muutoin tulokset esitetään oheisessa taulukossa. Taulukossa kulutus ilmoitetaan litroina sataa kilometriä kohti (l/100 km), teho hevosvoimina (hv), sylinteritilavuus kuutiosenttimetreinä (cm^3), massa kiloina (kg) ja kiihtyvyys nollasta sataan kilometriin tunnissa sekunteina (s).

Autojen polttoaineenkulutus, teho, moottorin tilavuus, omamassa ja kiihtyvyys

Merkki/malli	Kulutus	Teho	Tilavuus	Oma-massa	Kiihty-vyys
Audi A3 Sportback 1,8 TFSI	7,4	160	1 798	1 420	8,2
Citroën C5 2,0i VTR Berline	8,0	140	1 997	1 528	9,1
Citroën C6 3,0i V6 Exclusive	11,2	211	2 946	1 891	9,4
Fiat Albea 1,4 Sedan	6,9	77	1 368	1 105	14,0
Ford Focus 1,6 Ti-VCT Collection	6,6	116	1 596	1 275	10,8
Honda FR-V1,8i Executive AT	8,1	140	1 799	1 520	12,5
Hyundai Sonata 2,0 GL	7,8	145	1 998	1 473	10,5
Hyundai Sonata 3.3 V6	10,1	235	3 342	1 571	7,7
Kia Cee'd 1.6 EX	6,4	122	1 591	1 260	10,9
Kia Magentis 2,0 LX	7,7	144	1 998	1 440	10,2
Lexus LS	9,3	394	4 969	2 410	6,3
Opel Astra 1,6 Enjoy Edition	6,6	105	1 598	1 255	12,3
Peugeot 307 Premium Plus 1,6	7,4	109	1 587	1 245	11,7
Toyota Auris 1,6 Linea Sol	7,1	97	1 598	1 300	10,4
Toyota Corolla 1,4 VVT-i Linea Terra	6,7	97	1 398	1 265	12,6
Volkswagen Golf 1,6 Confortline	7,4	102	1 595	1 270	11,4
Volkswagen GrossGolf 1,4 TSI	7,3	140	1 390	1 457	9,3
Volkswagen Touran 1,4 TSI DSG	7,5	140	1 390	1 621	9,7
Volvo S80 3.2 Momentum A	7,9	238	3 192	1 656	7,9

Regressiomalli

Tarkoituksena on tehdä usean muuttujan regressioanalyysi siten, että selitettävänä muuttujana on polttoaineen kulutus. Selittävinä muuttujina ovat aluksi moottorin teho, sylinteritilavuus, auton omamassa ja kiihtyvyys nollasta sataan kilometriin tunnissa.

Aluksi kaikki lähtötiedot syötetään tilasto-ohjelmaan ja tehdään usean muuttujan lineaarinen regressioanalyysi. Tällöin saadaan seuraavan taulukon mukainen tuloste:

Coefficients[a]						
		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	3,90590	2,801		1,394	,185
	Teho (hv)	–,03851	,018	–2,298	–2,190	,046
	Moottorin tilavuus (cm^3)	,00234	,001	1,754	2,481	,026
	Omamassa (kg)	,00465	,002	1,111	2,400	,031
	Kiihtyvyys 0–100 km/h (s)	–,18582	,183	–,293	–1,018	,326

[a.] Dependent Variable: Kulutus (l/100 km)

Usean muuttujan regressioanalyysin tuloste.

Kun tarkastellaan selittävien muuttujien *t*-arvoja, voidaan havaita, että itseisarvoltaan pienin luku on auton kiihtyvyydellä. Koska *t*-arvo on itseisarvoltaan selvästi alle 2, auton kiihtyvyys poistetaan selittävien muuttujien joukosta. Näin jatkamalla voidaan poistaa selittävien muuttujien joukosta ne muuttujat, jotka todennäköisesti eivät vaikuta polttoaineen kulutukseen. Lopullinen tulos saadaan suoraan käyttämällä poistavaa valintaa, jolloin saadaan seuraava taulukko:

Coefficients[a]						
		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	3,90590	2,801		1,394	,185
	Teho (hv)	–,03851	,018	–2,298	–2,190	,046
	Moottorin tilavuus (cm^3)	,00234	,001	1,754	2,481	,026
	Omamassa (kg)	,00465	,002	1,111	2,400	,031
	Kiihtyvyys 0–100 km/h (s)	–,18582	,183	–,293	–1,018	,326
2	(Constant)	1,74700	1,833		,953	,356
	Teho (hv)	–,02888	,015	–1,723	–1,946	,071
	Moottorin tilavuus (cm^3)	,00198	,001	1,484	2,261	,039
	Omamassa (kg)	,00433	,002	1,033	2,260	,039

[a.] Dependent Variable: Kulutus (l/100 km)

Usean muuttujan regressioanalyysi poistavaa valintaa käyttämällä.

Taulukosta ilmenee, että jäljelle jääneiden selittävien muuttujien *t*-arvojen itseisarvo on noin kaksi, joten selittävien muuttujien poistaminen lopetetaan. Jäljelle jää regressiomalli, jossa on kolme selittävää muuttujaa eli moottorin teho, sylinteritilavuus ja auton omamassa:

Kulutus =
1,747 – 0,0289 · Teho + 0,00198 · Tilavuus + 0,00433 · Omamassa.

Mallin tarkastelua

Tarkastellaan lopullista mallia:

- On varsin helppo uskoa, ettei kiihtyvyydellä välttämättä olekaan kovin suurta vaikutusta polttoaineen kulutukseen, kun autolla ajetaan nopeusrajoitusten mukaisesti. Niinpä mallista poistui auton kiihtyvyys.
- Moottorin teho on vahvasti sidoksissa moottorin sylinteritilavuuteen, jolloin sylinteritilavuuteen sisältyy piilevänä tekijänä myös moottorin teho. Tämän voi tarkistaa esimerkiksi laskemalla sylinteritilavuuden ja tehon välisen korrelaatiokertoimen (0,969). Näistä muuttujista toisen olisi siten voinut jättää pois selittävien muuttujien joukosta jo alun perin.
- Moottorin tehoon liittyvä negatiivinen kerroin mallissa voi johtua siitä, että tehokkaat moottorit kuluttavat suhteellisesti vähemmän polttoainetta kuin pienitehoiset.
- Moottorin sylinteritilavuudella on mallissa positiivinen kerroin, sillä yleisesti ottaen suuremmat moottorit kuluttavat enemmän kuin pienet.
- Auton massan kasvattaminen lisää uskottavasti polttoaineenkulutusta.
- On kuitenkin muistettava, että jos olisi valittu jokin toinen joukko selittäviä muuttujia tai suurempi otos, malli voisi olla ainakin jossakin määrin toisenlainen.

Mallin hyvyys

Tarkastellaan aluksi selityskerrointa:

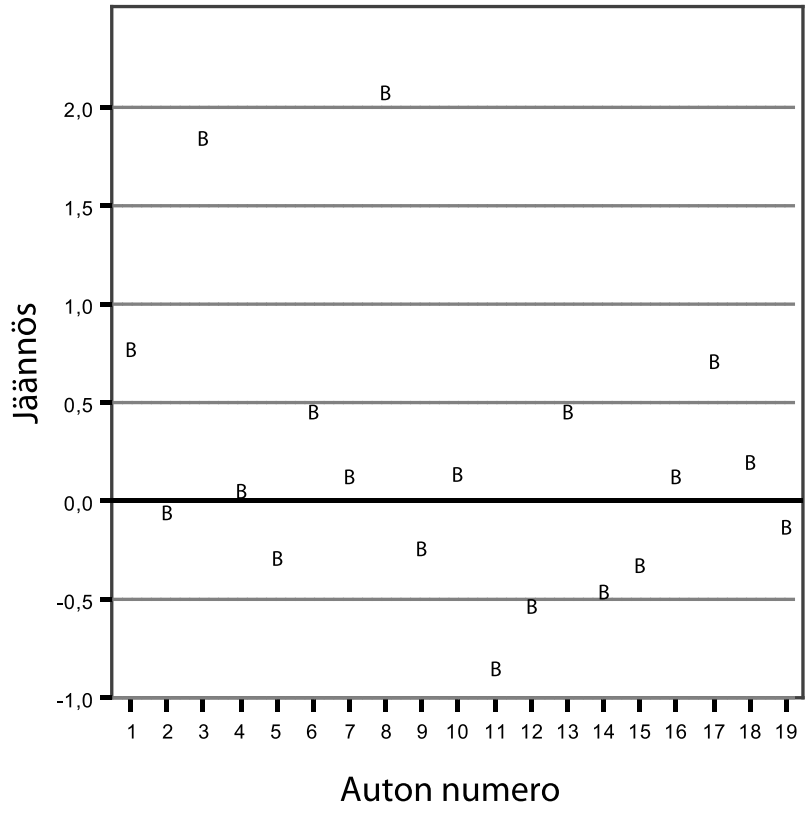
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,818[a]	,670	,575	,8039
2	,803[b]	,645	,574	,8049

[a.] Predictors: (Constant), Kiihtyvyys 0–100 km/h (s), Moottorin tilavuus (cm^3), omamassa (kg), Teho (hv)
[b.] Predictors: (Constant), Moottorin tilavuus (cm^3), Omamassa (kg), Teho (hv)

Mallien selityskertoimet.

Lopullisessa mallissa selityskerroin on edellä olevan taulukon mukaan 64,5 %. Mallin mukaan siis 64,5 % polttoaineen kulutuksen vaihtelusta voidaan selittää moottorin sylinteritilavuudella, moottorin teholla ja auton omalla massalla. Loput polttoaineen kulutuksen vaihtelusta johtuu muista tekijöistä. Mallin hyvyyttä tutkitaan vielä jäännöstermin e avulla. Tätä varten joudutaan määrittämään mallin mukaiset ennustearvot $\hat{y}_i$, joiden perusteella lasketaan jäännöstermin arvot $e_i = y_i - \hat{y}_i$. Jäännöstermin arvoista piirretään lopuksi hajontakuvio, jolloin saadaan seuraava kuva:



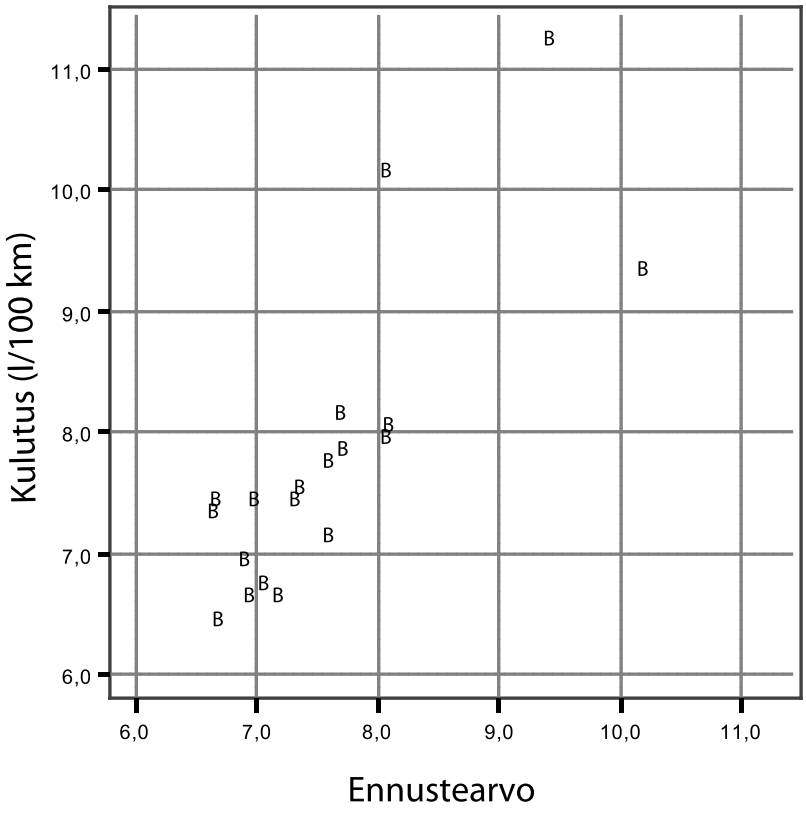
Jäännöstermin arvot hajontakuviona.

Kuvasta käy ilmi, että malli on melko hyvin määritelty, sillä jäännöstermin arvot ovat itseisarvoltaan melko pieniä ja ne ovat jakautuneet sopivasti nollan molemmille puolille, lukuun ottamatta Hyundai Sonata 3,3:a ja Citroën C6:tta. Kuvaajaa katsellessa tulee mieleen, että jäännöstermin arvot ikään kuin vaimenisivat vähitellen. Tässä ei kuitenkaan ole kyseessä aikasarja, joten autojen järjestyksellä ei ole merkitystä eikä kuviota voi siis katsoa jäännöstermin arvojen kehityskuviona.

Ennusteen luotettavuus

Ennusteen luotettavuutta voidaan myös arvioida piirtämällä hajontakuvio, jonka pisteet muodostuvat arvoista $(\hat{y}_i, y_i)$. Alla olevassa kuvassa esitetään kyseinen kuvaaja.

Kuvasta käy ilmi, että pisteparvi sijoittuu paria poikkeusta lukuun ottamatta melko hyvin kuvan kuvitellun halkaisijan molemmin puolin eikä suuria poikkeamia ole. Tällöin myös ennustearvot ovat todennäköisesti melko lähellä todellisia kulutusarvoja.



Ennustearvojen graafinen tarkastelu.

Ennusteita

Tutkitaan seuraavaksi, kuinka hyvin mallin perusteella laaditut ennusteet vastaavat todellisuutta. Sijoitetaan saatuun regressiomalliin

$$\text{Kulutus} = 1{,}747 - 0{,}0289 \cdot \text{Teho} + 0{,}00198 \cdot \text{Tilavuus} + 0{,}00433 \cdot \text{Omamassa}$$

kunkin auton teho, sylinterin tilavuus ja auton oma massa ja lasketaan ennustearvo. Esimerkiksi Audi A3:lle saadaan ennuste seuraavasti:

$$\text{Ennuste} = 1{,}747 - 0{,}0289 \cdot 160 + 0{,}00198 \cdot 1\,798 + 0{,}00433 \cdot 1\,420 = 6{,}8.$$

Autojen todelliset kulutuslukemat ja regressiomallin antamat ennusteet

Merkki/malli	Kulutus	Ennuste
Audi A3 Sportback 1,8 TFSI	7,4	6,8
Citroën C5 2,0i VTR Berline	8,0	8,3
Citroën C6 3,0i V6 Exclusive	11,2	9,7
Fiat Albea 1,4 Sedan	6,9	7,0
Ford Focus 1,6 Ti-VCT Collection	6,6	7,1
Honda FR-V1,8i Executive AT	8,1	7,8
Hyundai Sonata 2,0 GL	7,8	7,9
Hyundai Sonata 3.3 V6	10,1	8,4
Kia Cee'd 1.6 EX	6,4	6,8
Kia Magentis 2,0 LX	7,7	7,8
Lexus LS	9,3	10,6
Opel Astra 1,6 Enjoy Edition	6,6	7,3
Peugeot 307 Premium Plus 1,6	7,4	7,1
Toyota Auris 1,6 Linea Sol	7,1	7,7
Toyota Corolla 1,4 VVT-i Linea Terra	6,7	7,2
Volkswagen Golf 1,6 Confortline	7,4	7,5
Volkswagen GrossGolf 1,4 TSI	7,3	6,8
Volkswagen Touran 1,4 TSI DSG	7,5	7,5
Volvo S80 3.2 Momentum A	7,9	8,4

Tulkinta

Taulukossa olevat tiedot vastaavat vain uuden auton kulutusta ihanneoloissa. Todellisuudessa polttoaineenkulutukseen vaikuttavat monet muutkin tekijät, kuten auton ikä, kuljettajan ajotapa, edellisestä huollosta kulunut aika, säädöt, renkaat, tien kunto ja tyyppi, vuodenaika, muu liikenne ja nopeuden tasaisuus. Todennäköisesti siihen vaikuttavat myös auton ilmanvastuskerroin sekä vaihteiden lukumäärä ja käyttö, mutta niitä ei tässä selvityksessä päästy tutkimaan.

Jokaisella automerkillä kaupunki- ja maantieajon kulutuksissa on huomattavia eroja. Osa pienimmistä autoista on suunniteltu enemmän kaupunkiajoon, ja suuria autoja on tarkoitus käyttää enemmän maantieajoon. Tämän vuoksi kaupunkiajon ja maantieajon pelkkä yhdistetty kulutus ei ole täysin vertailukelpoinen eri tyyppisille autoille.

Edellä on saatu malli, jonka perusteella voidaan arvioida auton polttoaineen yhdistetty kulutus, kun tiedetään auton teho, sylinteritilavuus ja omamassa. Osoittautui, että mitä suurempi moottori ja painavampi auto, sitä suurempi on polttoaineen kulutus. Tulos sinällään ei yllätä ketään. Ratkaisevaa onkin, että nyt pystytään ennakoimaan polttoaineen kulutusta matemaattisen mallin avulla.

Saadut ennusteet osoittautuivat kohtuullisen tarkoiksi, jälleen samoja kahta poikkeusta lukuun ottamatta. Selityskertoimen perusteella kovin paljon tarkempaa tulosta ei olisi edes osattu odottaakaan.

Jäännöstermien kuviota tarkasteltaessa havaitaan, että kahden auton tulokset poikkeavat selvästi muista. Lähempi tarkastelu osoitti, että muista poikkeavat arvot liittyvät autoihin, joiden sylinteritilavuus on suuri. Olisikin ollut järkevää jakaa autot moottoritilavuudeltaan kahteen ryhmään, esimerkiksi korkeintaan kaksilitraisiin ja yli kaksilitraisiin (ks. tehtävä 14).

TEHTÄVIÄ

1. Mitkä seuraavista muuttujapareista muodostavat tilastollisen ja mitkä funktionaalisen yhteyden:
- **a)** henkilön pituus ja paino
- **b)** neliön sivu ja neliön pinta-ala
- **c)** tuotteen mainontaan käytetty rahamäärä ja tuotteen myynti
- **d)** ilmanpaine ja korkeus merenpinnasta?
- **e)** ikä ja verenpaine?

2. Oletetaan, että asumiskulut riippuvat jollakin tavalla käytettävissä olevista nettotuloista. Tämän perusteella kerättiin seuraava havaintoaineisto:

Tulot (100 €)	23,4	25,2	27,3	29,6	31,4	35,4	36,5	38,1	40,7	44,8
Asumiskulut (100 €)	6,3	5,7	7,2	9,9	13,2	10,8	12,0	17,1	18,4	22,6

- **a)** Tutki, onko lineaarinen malli mahdollinen. Jos vastaus on myönteinen, piirrä silmämääräisesti viivoitinta käyttäen pistejoukkoa kuvaava suora.
- **b)** Laadi graafisesti ennuste asumiskuluista, kun käytettävissä olevat tulot ovat 3 000 euroa.
- **c)** Määritä lineaarisen mallin parametrien arvot ja laadi mallin avulla samanlainen ennuste kuin edellä. Vertaa näin saatuja ennusteita.

3. Arveltiin, että bruttokansantuote (USD/henkilö) vaikuttaa hiilidioksidipäästöjen määrään (tn/henkilö). Tutki oheisen tilastoaineiston hajontakuviota ja määritä regressiosuoran yhtälö. Mikä on johtopäätös? Arvioi hiilidioksidipäästöjen määrä, kun bruttokansantuote on 25 000 USD/henkilö. (Lähde: Tilastokeskus.)

Maa	**Bkt (USD)**	**CO_2-päästöt (tn/henkilö)**	**Maa**	**Bkt (USD)**	**CO_2-päästöt (tn/henkilö)**
Australia	34 932	17,5	Kiina	1 716	3,7
Brasilia	4 789	1,8	Ranska	33 925	6,2
Iso-Britannia	37 042	9,0	Ruotsi	39 658	5,8
Espanja	25 997	7,7	Saksa	33 865	10,3
Etelä-Afrikka	5 160	7,6	Suomi	37 320	13,2
Intia	712	1,0	Tanska	47 906	9,4
Italia	30 525	8,0	Venäjä	5 323	10,6
Kanada	35 105	17,2	Yhdysvallat	41 960	19,7

4. Oheisessa taulukossa on eräiden yritysten pörssikurssit ja niiden jakamat osingot euroina.

Yritys	Pörssikurssi	Osinko	Yritys	Pörssikurssi	Osinko
Elisa	21,37	1,50	Nordea	12,34	0,49
Finnair	12,52	0,10	Olvi	27,49	0,65
Fortum	24,38	1,26	Orion	18,20	1,00
Kemira	16,15	0,48	Outokumpu	26,47	1,10
Kesko	49,95	1,50	Sampo	23,30	1,20
Metso	40,38	1,50	SanomaWSOY	23,60	0,95
Neste Oil	27,47	0,90	UPM-Kymmene	19,16	0,75
Nokia	19,74	0,43			

a) Tutki, onko lineaarinen malli mahdollinen, kun pörssikurssilla selitetään osingon määrää? Jos vastaus on myönteinen, piirrä silmämääräisesti viivoitinta käyttäen pistejoukkoa kuvaava suora.
b) Määritä lineaarisen mallin parametrien arvot ja laadi mallin avulla ennuste osingolle, kun pörssikurssi on 30 euroa. Tarkista tulos kuvasta.
c) Määritä selityskerroin. Mitä siitä voi päätellä?
d) Tutki, miten selityskerroin ja ennuste muuttuvat, jos Elisa ja Finnair poistetaan tilastoaineistosta.

5. Tutkittiin ratakiskon lämpölaajenemista. Asian selvittämiseksi mitattiin 30 metrin ratakiskon pituus eri lämpötiloissa. Tulokset olivat seuraavat:

Pituus (mm)	30 003	30 008	30 014	30 018	30 023	30 025	30 028	30 031
Lämpötila (°C)	–42	–28	–10	0	12	28	32	40

a) Piirrä tilastoaineistosta hajontakuvio.
b) Määritä ratakiskon pituuden ja lämpötilan välinen regressiosuoran yhtälö.
c) Tulkitse regressiosuoran yhtälön vakiotermi b_0 käytännön kannalta.
d) Tampereen ja Helsingin välinen matka rautateitä pitkin on ilmoituksen mukaan 187 km. Kuinka monta metriä pitempi on Tampereen ja Helsingin välisten yhden raiteen ratakiskojen yhteenlaskettu pituus kesähelteellä (lämpötila 30 °C) kuin talvipakkasella (lämpötila –35 °C)?

6. Laadi ennuste elinajanodotteesta esimerkin 1 tilastoaineiston perusteella, kun bruttokansantuote on 100 000 USD. Mikä on johtopäätös?

7. Sanomalehdestä seurattiin Helsingin seudun kerrostaloasuntojen koon vaikutusta asunnon myyntihintaan. Tulokset olivat seuraavat:

Koko (m^3)	Hinta (1 000 €)	Koko (m^3)	Hinta (1 000 €)
25	108	70	173
34	172	79	290
40	186	87	474
41	198	92	340
45	208	106	387
59	285	125	497
61	144	146	530
69	295		

a) Piirrä hajontakuvio.
b) Määritä regressiomalli ja laadi ennuste asunnon hinnasta, kun asunnon koko on 50 m^2.
c) Minkä suuruisen kerrostaloasunnon todennäköisesti saa 350 000 eurolla?
d) Määritä selityskerroin ja tulkitse se.
e) Minkä kokoisten kerrostaloasuntojen hintaa on järkevää tarkastella kohdassa b) saadun regressiomallin avulla?

8. Määritä esimerkin 3 tilastoaineistosta regressiosuoran yhtälö ja vertaa tulosta selityskertoimen perusteella esimerkin tulokseen. Mikä on vertailun johtopäätös?

9. Kuvaa tietokoneohjelman avulla esimerkin 1 tilastoaineistoa toisen asteen käyrän avulla. Määritä selityskerroin ja tutki, onko saatu malli parempi kuin esimerkissä 1 saatu malli.

10. Laadi esimerkin 6 perusteella ennuste, kun bruttokansantuote on 39 300 USD ja imeväiskuolleisuus on 3,6. Vastaako tulos Suomen tietoja?

11. Muodosta toisen asteen malli esimerkissä 3 esitetylle maapallon väestönkehitykselle.
a) Laadi vuoden 2010 ennuste.
b) Muodosta jäännöstermin hajontakuvio ja vertaa sitä sivujen 285 ja 286 hajontakuvioihin.
c) Testaa malleja laatimalla vuoden 1950 ennuste eksponenttifunktion, regressiosuoran ja toisen asteen mallin avulla.

12. Tutkittiin tietynmerkkisen auton hintaan vaikuttavia tekijöitä. Asian selvittämiseksi kerättiin seuraava tilasto. Tilastossa hinta on ilmoitettu tuhansina euroina, ikä täysinä vuosina ja ajettu kilometrimäärä tuhansina kilometreinä.

Hinta (1 000 €)	Ikä (vuosia)	Ajettu (1 000 km)	Hinta (1 000 €)	Ikä (vuosia)	Ajettu (1 000 km)
18,9	5	97	29,9	7	124
31,9	4	45	29,9	5	98
18,5	3	83	35,9	3	91
33,9	1	17	21,9	3	46
24,9	4	143	16,9	9	124
15,2	8	189	8,9	14	199
26,9	3	82	24,9	4	74
26,9	6	69			

Määritä auton hintaa kuvaava matemaattinen malli niin, että käytät selittävinä muuttujina

a) pelkästään auton ikää,
b) pelkästään ajettua matkaa ja
c) sekä auton ikää että ajettua matkaa.
d) Määritä auton hinta, kun sen ikä on 3 vuotta ja sillä on ajettu 80 000 km.
e) Piirrä a)-kohdan osalta jäännöstermin hajontakuvio.

13. Laadi tilasto-ohjelman avulla ennusteen luotettavuutta kuvaavat 95 %:n ennustevyöhykkeet, kun tilastoaineistona on

a) tehtävä 2,
b) tehtävä 3.
c) Määritä kuvion perusteella molempien tapausten ennusteväli, kun ennusteväli määritetään selittävän muuttujan keskiarvon kohdalta.

14. Case C:n jäännöstermien kuviota tarkasteltaessa havaittiin, että kahden auton tulokset poikkeavat selvästi muista. Lähempi tarkastelu osoitti, että muista poikkeavat arvot liittyvät autoihin, joiden sylinteritilavuus on suuri. Olisikin ollut järkevää jakaa autot moottoritilavuudeltaan kahteen ryhmään, esimerkiksi korkeintaan kaksilitraisiin ja yli kaksilitraisiin.

Seuraavaksi tarkastellaan Case C:tä siten, että selvitykseen otetaan mukaan vain ne autot, joiden sylinteritilavuus on korkeintaan kaksi litraa.

a) Muodosta usean muuttujan regressiomalli polttoaineenkulutuksesta, kun selittävinä muuttujina ovat mahdollisesti moottorin teho ja tilavuus sekä auton omamassa ja kiihtyvyys.
b) Määritä selityskerroin.
c) Muodosta jäännöstermien hajontakuvio
d) Laadi mallin avulla kutakin automerkkiä ja -mallia koskevat ennusteet ja vertaa tuloksia ilmoitettuihin arvoihin.
e) Vertaa ennusteita Case C:n tuloksiin.
f) Mieti, oliko sylinteritilavuutta koskeva rajaus järkevä tehdä.

10
Aikasarjat

Ympärillämme tapahtuu koko ajan paljon elämäämme vaikuttavia asioita, kuten asuntotuotantoa, energiankulutusta ja ympäristön saastumista. Kaikkia näitä voimme seurata ajan kuluessa, ja jos pidämme havainnoistamme kirjaa, syntyy aikasarja. Jos vielä kuvaamme aikasarjaa graafisesti, voimme ainakin silmämääräisesti nähdä kehityksen suunnan ja mahdollisesti ennakoida tulevaisuutta. Tässä luvussa tutustutaan menetelmiin, joilla tutkitaan aikasarjoja ja laaditaan ennusteita.

10.1 Kuvaajat ja komponentit

Aikasarjojen tutkiminen on osoittautunut mielenkiintoiseksi, joskin vaativaksi tehtäväksi. Varsinkin talouteen liittyvien aikasarjojen seuranta on monille yrityksille suorastaan välttämätöntä, sillä taloudelliset päätökset pohjautuvat tiettyihin olettamuksiin. Yksi tärkeimmistä olettamuksista on se, että menneisyydessä havaittu säännönmukaisuus usein toistuu jonkinasteisena myös tulevaisuudessa. Aikasarja-analyysin keskeisimpiä tehtäviä onkin mahdollisen säännönmukaisuuden löytäminen ja ennusteiden tekeminen.

Aikasarjan määritelmä
Aikasarja on ajan suhteen järjestetty havaintoarvojen joukko.

Aikasarja voi olla joko **jatkuva** tai **diskreetti** eli epäjatkuva. Diskreetin aikasarjan havaintoarvot on otettu vain tiettyinä, tavallisesti tasavälisinä ajankohtina. Jatkuvaan aikasarjaan päädytään esimerkiksi silloin, kun havaintoarvoja mitataan ja tallennetaan jatkuvasti. Tällöin kyseessä on yleensä automaattinen mittaus, esimerkiksi jonkin teollisen prosessin valvonta.

Seuraavassa käsitellään vain diskreettejä aikasarjoja, jolloin aikasarjan arvoja merkitään y_1:llä, y_2:llä,... y_n:llä. Merkintä on looginen siinä mielessä, että y:llä on jo aikaisemmin merkitty selitettävää muuttujaa. Toisaalta selittävänä muuttujana on aika, jota aikasarjojen yhteydessä yleisesti merkitään t:llä. Aikasarjaa voidaan siten yleisesti merkitä y_t:llä, jossa t:n ajatellaan saavan arvot 1, 2,... n.

Aikasarjan arvot eivät yleensä ole toisistaan riippumattomia, joten regressioanalyysin menetelmiä ei voida välttämättä käyttää sellaisenaan, sillä havaintoarvojen järjestys on otettava huomioon. Tällöin voidaan joutua käyttämään nimenomaan aikasarjojen tutkimiseen kehitettyjä menetelmiä.

Aikasarjoja analysoidaan nykyään lähes yksinomaan tietokoneohjelmien avulla. Ohjelmat poikkeavat toisistaan jonkin verran, mutta useimpiin ohjelmiin sisältyy ainakin joitakin aikasarjoihin liittyviä osia. Tämän luvun esimerkkien ratkaisemisessa käytetäänkin pääasiassa tietokonetta.

Aikasarjan kuvaajat

Aikaisemmissa luvuissa on huomattu tärkeäksi kuvata havaintojoukkoa graafisesti. Sama periaate sopii erityisen hyvin myös aikasarjoihin, joten paras tapa aloittaa aikasarjan tutkiminen on sen graafinen kuvaaminen. Tavaksi on tullut valita aika vaaka-akseliksi. Itse aikasarjan arvot haetaan pystyakselilta, joten aikasarjan mahdolliset arvot määräävät pystyakselin jaotuksen. Akselit voidaan tarpeen mukaan katkaista origon ympäristöstä, mutta kuitenkin siten, että kaikki samaan aikasarjaan kuuluvat pisteet tulevat aina katkaisun samalle puolelle.

Jos aikasarja muodostuu yksittäisistä, esimerkiksi tiettyinä ajankohtina mitatuista arvoista, aikasarjaa kuvaava piste merkitään vastaavan ajankohdan kohdalle. Jos taas aikasarjan arvot kuvaavat tietyn ajanjakson, esimerkiksi koko vuoden, arvoa, aikasarjaa kuvaava piste merkitään joko ajanjakson keskikohdan tai ajanjakson päättymistä kuvaavan jakoviivan kohdalle.

Aikasarjan kuvaaja voidaan tehdä monella tavalla. Kaikkein tavallisin on tapa, jossa aikasarja kuvataan sellaisenaan eli pisteinä (t, y_t). Peräkkäiset aikasarjan pisteet yhdistetään tämän jälkeen toisiinsa (suorilla) viivoilla, jolloin kuvaajasta tulee yhtenäinen ja mahdollinen sään-

nönmukaisuus on helppo hahmottaa. Joissakin tapauksissa aikasarjan arvoja joudutaan aluksi muuntamaan. Tällöin kyseeseen tulevat muun muassa vaihtoehdot $(t, \log y_t)$, $(t, \sqrt{y_t})$, $(t, y_t - y_{t-1})$ ja $(t, y_t/y_{t-1})$. Logaritmi- ja neliöjuurimuunnoksella pyritään suoristamaan aikasarjan kuvaaja, jolloin muunnoksen jälkeen voidaan soveltaa lineaarista mallia. $y_t - y_{t-1}$ kuvaa absoluuttista muutosta verrattuna edelliseen arvoon. y_t/y_{t-1} taas kuvaa suhteellista muutosta verrattuna edelliseen arvoon.

Aikasarjojen muunnoksiin kuuluu myös **indeksien** käyttö, sillä indeksien avulla eri aikasarjojen vertailu helpottuu. Aikasarjan muuntaminen indekseiksi tapahtuu valitsemalla jokin aikasarjan arvoista perusarvoksi ja merkitsemällä sitä luvulla 100. Muut aikasarjan arvot lasketaan sitten prosentteina perusarvosta.

Jos aikasarjan luvut on ilmoitettu rahana, voidaan inflaation vaikutus ottaa huomioon. Tämä tarkoittaa sitä, että aikasarjan luvut aluksi deflatoidaan ja tarkastelu kohdistetaan tähän deflatoituun aikasarjaan. Aikasarja **deflatoidaan** siten, että kunkin ajankohdan arvo jaetaan vastaavan ajankohdan indeksillä ja osamäärä kerrotaan vertailuajankohdan indeksillä.

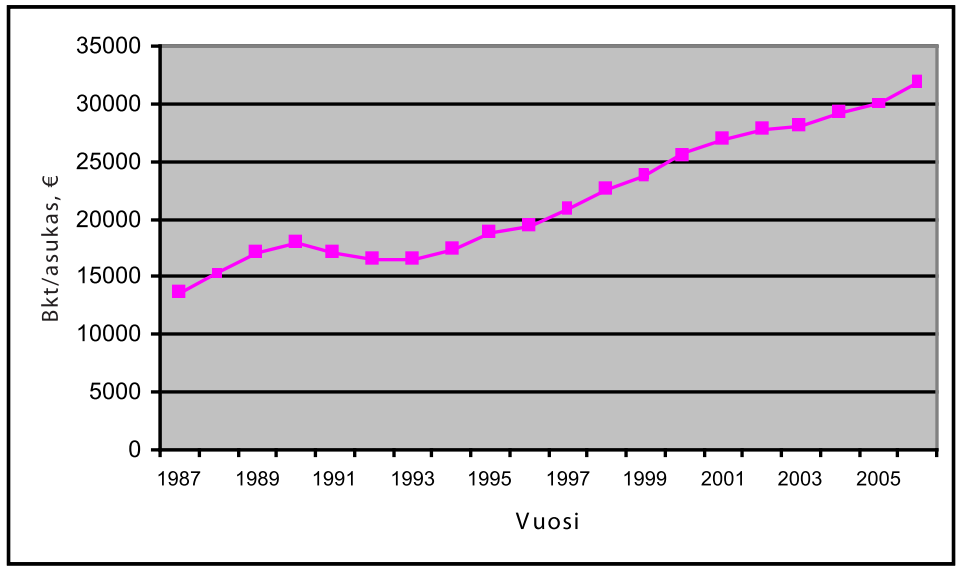
Kuvaaja laaditaan tavallisesti tietokoneohjelmalla. Tämä mahdollistaa kuvan monipuolisen muokkauksen.

ESIMERKKI 1

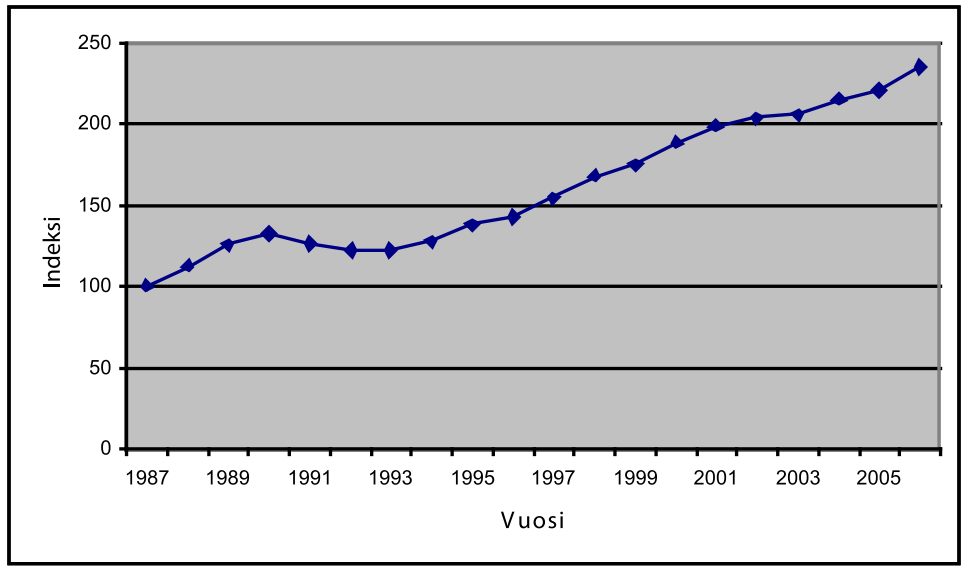
Bruttokansantuote asukasta kohti (Bkt/asukas, €) vuosina 1987–2006 oli Suomessa markkinahintaan seuraava (lähde: Tilastokeskus):

Vuosi	Bkt/asukas, €	Vuosi	Bkt/asukas, €
1987	13 547	1997	20 939
1988	15 308	1998	22 727
1989	17 079	1999	23 765
1990	18 000	2000	25 555
1991	17 092	2001	26 960
1992	16 470	2002	27 682
1993	16 566	2003	27 995
1994	17 312	2004	29 144
1995	18 778	2005	29 964
1996	19 367	2006	31 886

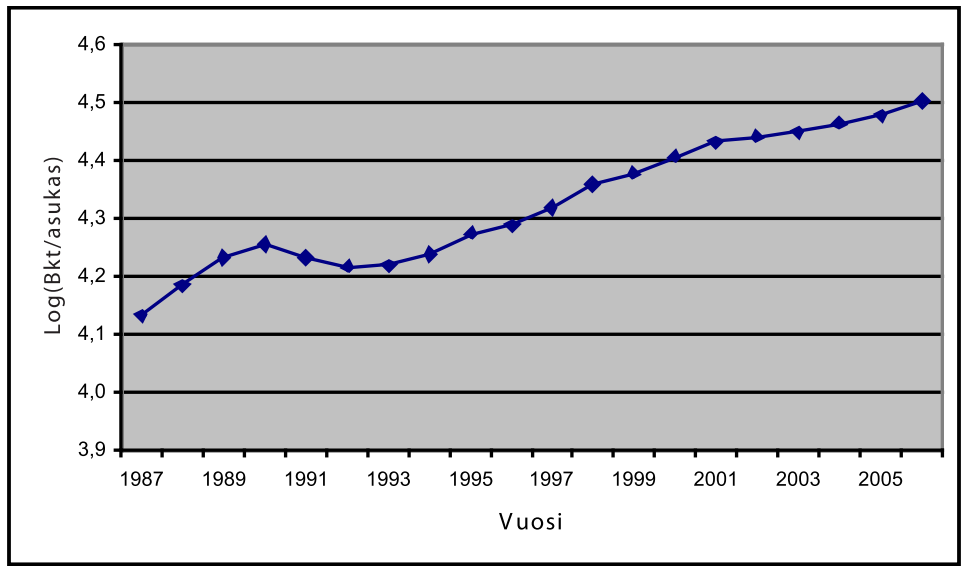
Edellisellä sivulla olevaa aikasarjaa voidaan kuvata eri vaihtoehtoja käyttämällä seuraavasti:



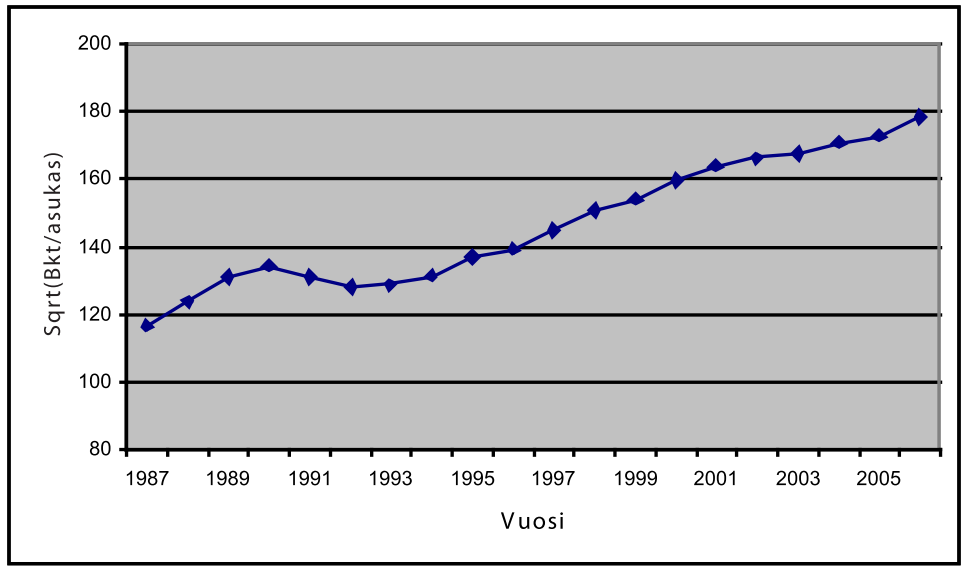
Bruttokansantuote (Bkt) markkinahintaan 1987–2006.



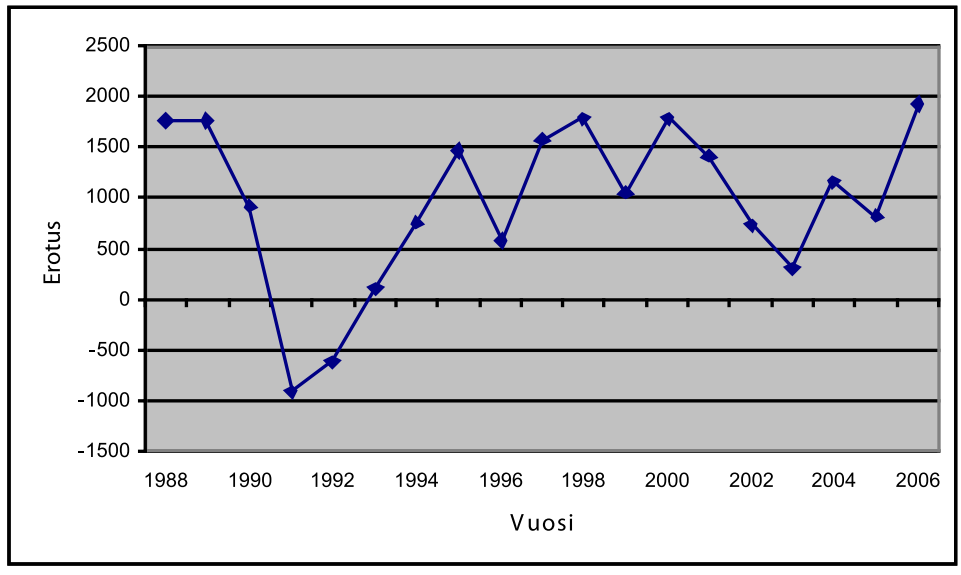
Bruttokansantuote (Bkt) markkinahintaan 1987–2006 indekseinä.



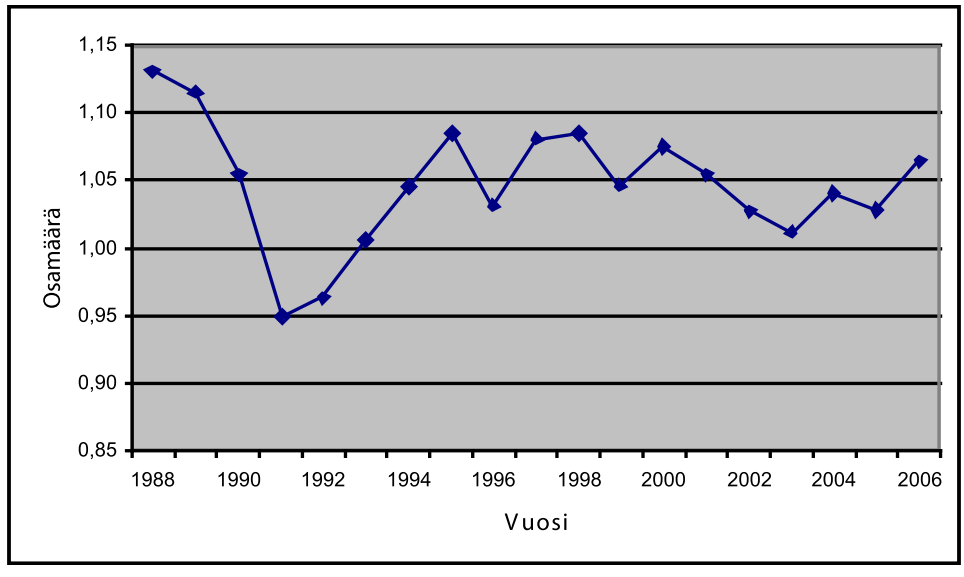
Bruttokansantuote (Bkt) markkinahintaan 1987–2006 logaritmimuunnoksena.



Bruttokansantuote (Bkt) markkinahintaan 1987–2006 neliöjuurimuunnoksena.



Bruttokansantuote (Bkt) markkinahintaan 1987–2006 erotuksena $y_t - y_{t-1}$.



Bruttokansantuote (Bkt) markkinahintaan 1987–2006 osamääränä y_t / y_{t-1}.

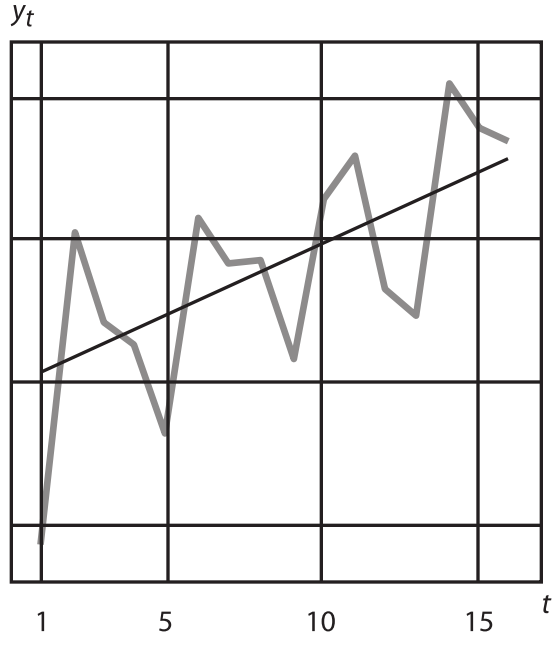
Aikasarjan komponentit

Klassinen aikasarja-analyysi lähtee olettamuksesta, että havaittua aikasarjaa voidaan kuvata yhdellä tai useammalla komponentilla. Tällöin erotetaan kolme pääkomponenttia eli trendi, kausivaihtelu ja suhdannevaihtelu. Ellei aikasarjan arvoja voida täysin selittää näiden komponenttien avulla, kuten useimmissa tapauksissa käy, loppu jää satunnaisvaihtelukomponentin tiliin. On kuitenkin huomattava, että kaikissa aikasarjoissa ei läheskään aina ole kaikkia edellä lueteltuja komponentteja.

Trendi (T)

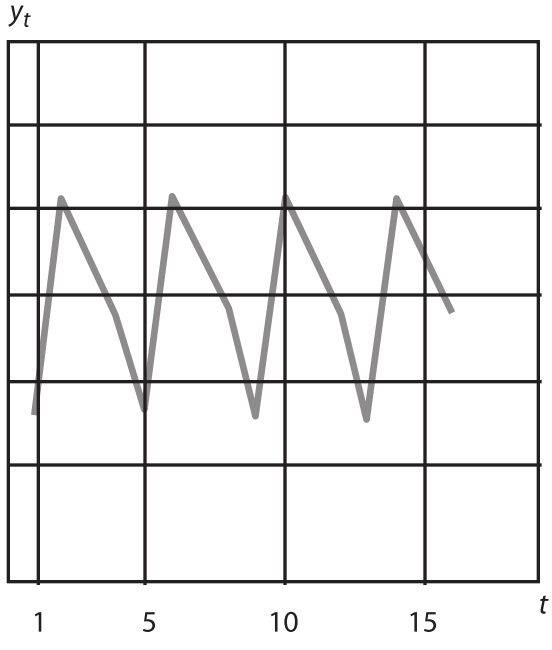
Trendi (trend) osoittaa pitkän aikavälin kehityssuunnan. Tavallisesti trendin määritelmään kuuluu kaksi ehtoa: kehityssuunnan pitää olla pitkäkestoista, ja sen pitää kulkea vain yhteen suuntaan, siis joko ylöspäin tai alaspäin. Trendin ei kuitenkaan tarvitse olla suoraviivaista.

Trendin syntyyn vaikuttavat muun muassa väestönkasvu, teknologian kehittyminen, inflaatio sekä ympäristö- ja käyttäytymismuutokset.



Kausivaihtelu (S)

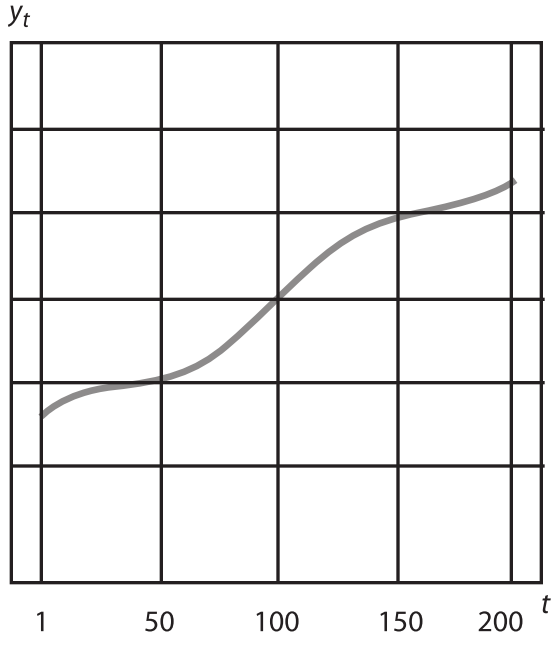
Kausivaihtelu (seasonal fluctuation) on jaksottaista, melko säännöllistä heilahtelua trendin molemmin puolin siten, että jakson pituus on korkeintaan yksi vuosi. Vuodenajat aiheuttavat eniten kausivaihtelua, mutta kausivaihteluksi voidaan tulkita esimerkiksi lämpötilan vaihtelut vuorokausittain. Kausivaihtelua esiintyy erityisen paljon talouteen liittyvissä aikasarjoissa.



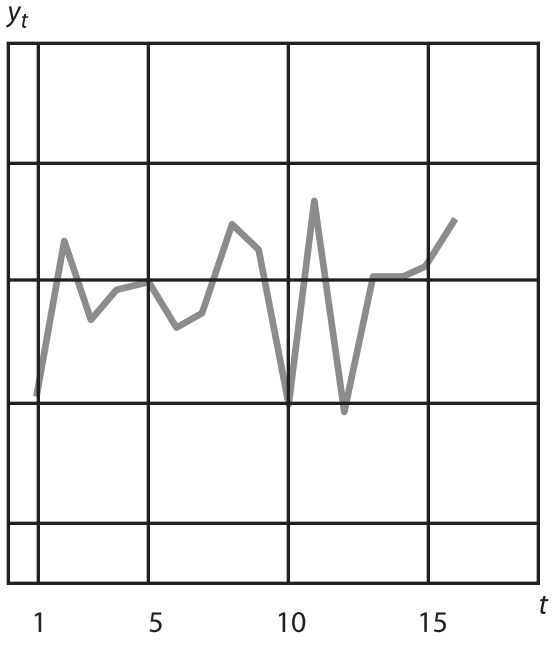
Suhdannevaihtelu (C)

Suhdannevaihtelu (cyclical behavior) on pitkän aikavälin heilahtelua. Keskeistä suhdannevaihtelussa on se, että jakson pituus on yli vuosi. Suhdannevaihtelu ei ole niin säännöllistä kuin kausivaihtelu, sillä suhdannevaihtelun aallonpituus ja heilahteluväli tavallisesti vaihtelee jaksosta toiseen.

Usein trendiä ja suhdannevaihtelua on vaikea erottaa toisistaan. Tämä seikka tulee esille varsinkin silloin, kun suhdannevaihtelun jakso on pitkä.



Satunnaisvaihtelu (I)
Satunnaisvaihtelu eli epäsäännöllinen vaihtelu (irregular movement) kuvaa sitä osaa aikasarjasta, jota ei voida selittää trendin, kausi- ja suhdannevaihtelun avulla. Satunnaisvaihtelua aiheuttavat muun muassa säätilan äkilliset muutokset, lakot, yllättävät poliittiset tapahtumat ja poikkeukselliset talousuutiset.



Komponenttien yhdistäminen

Klassisessa aikasarja-analyysissä oletetaan, että alkuperäinen aikasarja voidaan jakaa komponentteihin. Jako tehdään sen vuoksi, että kutakin komponenttia on sinänsä helpompi tutkia erikseen kuin koko alkuperäistä aikasarjaa. Tällöin komponentin käyttäytyminen ei sisällä niin monta muuta häiriötekijää ja sen mahdollinen säännönmukaisuus tulee paremmin esille.

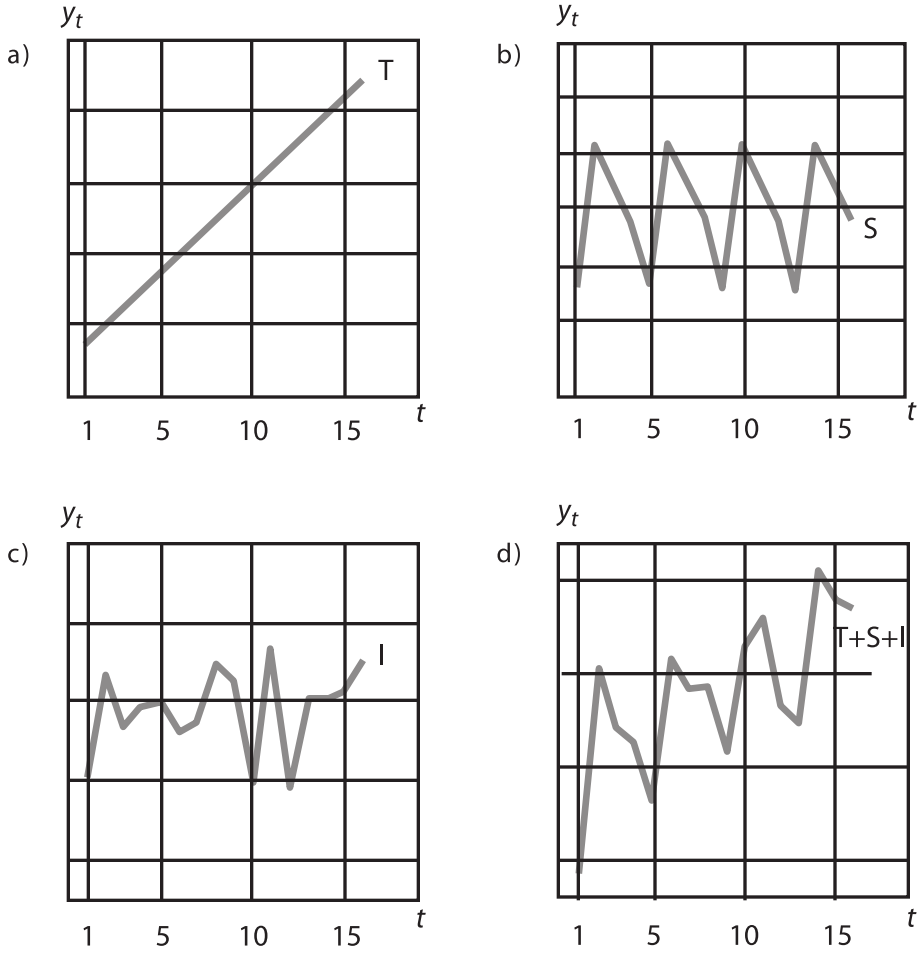
Kun komponentit on saatu arvioiduiksi mahdollisimman tarkasti, ne yhdistetään. Komponenteista muodostetaan malli yleensä jommallakummalla seuraavista tavoista:

$Y = T + S + C + I$ (summamalli)
$Y = T \cdot S \cdot C \cdot I$ (tulomalli).

Malli valitaan sen mukaan, kumpi niistä kuvaa paremmin alkuperäistä aikasarjaa. Asian tutkimiseksi muodostetaan mallin avulla aikasarjaa kuvaavat estimaatit, joita sitten verrataan alkuperäisen aikasarjan arvoihin. Malli on sitä parempi, mitä lähempänä mallista lasketut estimaatit ovat alkuperäisen aikasarjan arvoja. Mallin hyvyyttä voidaan tutkia samalla tavalla kuin regressiomallia.

ESIMERKKI 2

Tarkastellaan seuraavaa keinotekoista esimerkkiä, josta käyvät ilmi komponentit ja niistä muodostettu aikasarja. Esimerkissä käytetään summamallia.



Aikasarjan muodostuminen komponenteista.

10.2 Trendi

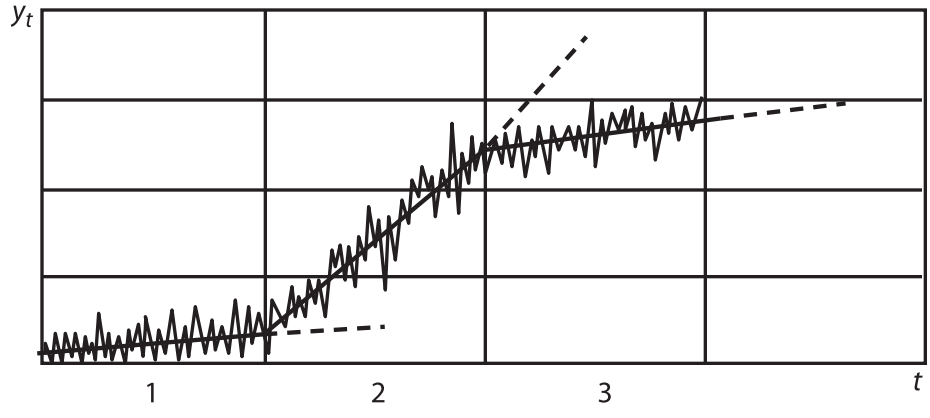
Aikasarjaa tutkittaessa komponentit otetaan huomioon kahdella tavalla. Yhtäältä pyritään tutkimaan kukin komponentti, sen käyttäytyminen ja mahdollinen matemaattinen malli ja toisaalta pyritään eliminoimaan kyseisen komponentin vaikutus tarkasteltavasta aikasarjasta.

Trendin määrittäminen

Trendiä kuvaava malli saattaa muuttua paljonkin sen perusteella, millaista aikajaksoa tarkastellaan. Tilannetta vaikeuttaa usein se, ettei suhdannevaihtelua pystytä erottamaan aikasarjasta vaan säännönmukaiselta vaikuttava pitkäaikainen muutos menee kokonaan trendin tiliin.

ESIMERKKI 3

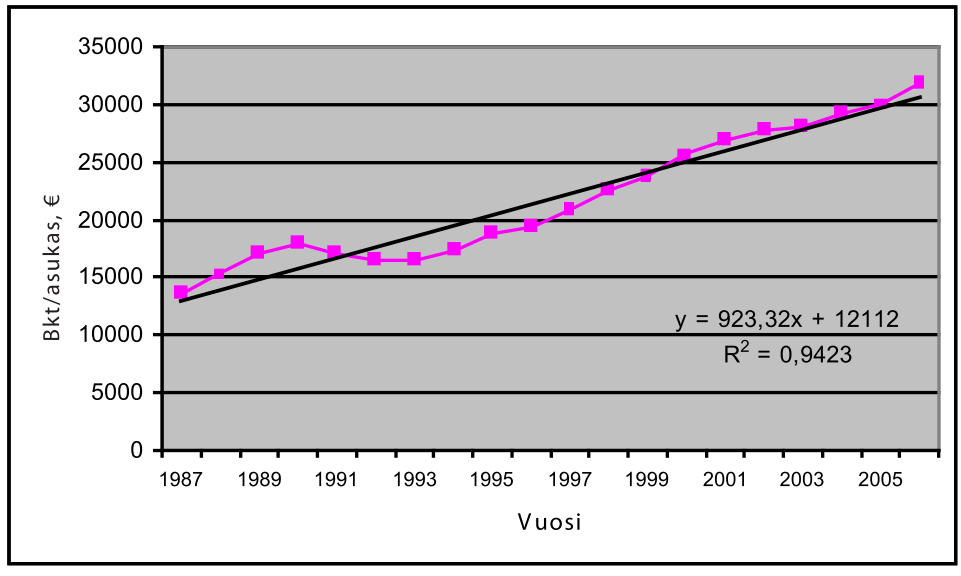
Trendi kuvaa aikasarjan yleistä kehityssuuntaa tietyllä aikavälillä. Jos kehitys jatkuu samanlaisena, trendiä voidaan käyttää ennusteiden laatimiseen. Kehitys voi kuitenkin muuttua, vaikkei sitä ehkä heti havaitakaan. Kuviossa pyritään selvittämään yksinkertaisesti, ettei yhdellä aikavälillä määritetty trendi kuvaakaan seuraavalla aikavälillä esiintyviä aikasarjan arvoja. Näin trendiin vaikuttaa suuresti se, miten pitkältä aikaväliltä se määritetään. Tässäkin osavälien trendit poikkeavat keskenään ja myös koko aikasarjan yhteisestä trendistä. Edellisestä voidaan havaita, että ennusteiden laatiminen pelkästään trendiä käyttämällä sisältää riskejä. Ennusteita laadittaessa on siis otettava huomioon myös muut komponentit.



Trendi pyrkii tasoittamaan aikasarjassa esiintyvät heilahtelut. Sama tasoitusominaisuus on pienimmän neliösumman menetelmällä, joten myös sillä määritetyt regressiomallit soveltuvat usein sellaisenaan aikasarjan trendin määrittämiseen.

ESIMERKKI 4

Määritä esimerkissä 1 olevan aikasarjan lineaarisen trendin yhtälö. Määritetään Excelillä trendisuoran yhtälö, kun vuotta 1987 merkitään *t*:n arvolla yksi. Tietokoneohjelmista saa tavallisesti samalla kertaa aikasarjan kuvaajan, trendisuoran kuvaajan ja trendisuoran yhtälön.



Bruttokansantuote (Bkt) markkinahintaan 1987–2006.

Kuvasta havaitaan, että lineaarisen trendin yhtälö on

$y = 12\,112 + 923{,}32 \cdot t$ ($t = 1$ vuonna 1987).

Excel antaa seuraavan taulukon, josta saadaan trendisuoran yhtälö:

	t					
	Kertoimet	**Keski-virhe**	**Tunnus-luvut**	***P*-arvo**	**Alin 95 %**	**Ylin 95 %**
Leikkaus-piste	12111,98	645,01	18,78	0,00	10756,87	13467,10
Vuosi	923,32	53,84	17,15	0,00	810,19	1036,44

Jottei tarvitsisi laskea, kuinka monta vuotta on kulunut vuodesta 1987 alkaen esimerkiksi vuoteen 2010, yhtälö kirjoitetaan muotoon

$y = 12\,112 + 923{,}32 \cdot (t - 1986)$ (t = tarkasteltava vuosiluku).

Lineaarisen trendin olemassaolon testaus

Aina ei ole varmaa, että aikasarja sisältää trendin. Asia voidaan testata esimerkiksi Spearmanin järjestyskorrelaatiokertoimen avulla, jolloin testiä sanotaan **Danielsin testiksi**. Testiä varten lasketaan Spearmanin järjestyskorrelaatiokerroin muuttujista t ja y_t. Jos korrelaatiokertoimen itseisarvo on suurempi kuin kirjan liitteenä olevan taulukon kriittinen arvo, aikasarjassa on todennäköisesti trendi. Trendi on nouseva, jos $r_S > 0$, ja laskeva, jos $r_S < 0$.

Lineaarisen trendin testaamiseksi voidaan käyttää Pearsonin korrelaatiokerrointa. Testimuuttujana käytetään nyt lauseketta

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

joka noudattaa t-jakaumaa vapausasteilla $n-2$. Jos aikasarjasta laskettu t:n itseisarvo on suurempi kuin t-jakauman kriittinen arvo, aikasarjassa on todennäköisesti lineaarinen trendi.

ESIMERKKI 5

Testataan, onko esimerkin 1 tilastoaineistosta löydettävissä lineaarinen trendi.

Määritetään aluksi korrelaatiokerroin (esimerkiksi kuvasta sivulla 317), $r = 0{,}9707$:

$$t = \frac{0{,}9707 \cdot \sqrt{20-2}}{\sqrt{1-0{,}9707^2}} = 17{,}14.$$

Tulkinta: Koska t:n itseisarvo ylittää selvästi kaikki t-jakauman taulukossa olevat kriittiset arvot, testin perusteella aikasarjaan todennäköisesti sisältyy lineaarinen nouseva trendi.

10.3 Kausivaihtelu

Kausivaihtelua aiheuttavat varsinkin säähän ja lämpötilaan sekä vuodenaikaan liittyvät seikat. Jos uskotaan, että aikasarjassa on kausivaihtelua, ensimmäiseksi on löydettävä kausivaihtelujakson pituus L eli se, kuinka monta aikasarjan arvoa kuuluu aina yhteen jaksoon. Tämä pyritään selvittämään joko tutkimalla aikasarjan arvoja tai tarkastelemalla aikasarjan kuvaajaa. Kuvaajassa jakson pituus on kahden peräkkäisen huipun tai kahden peräkkäisen aallonpohjan välinen matka.

Kausivaihtelun olemassaolo

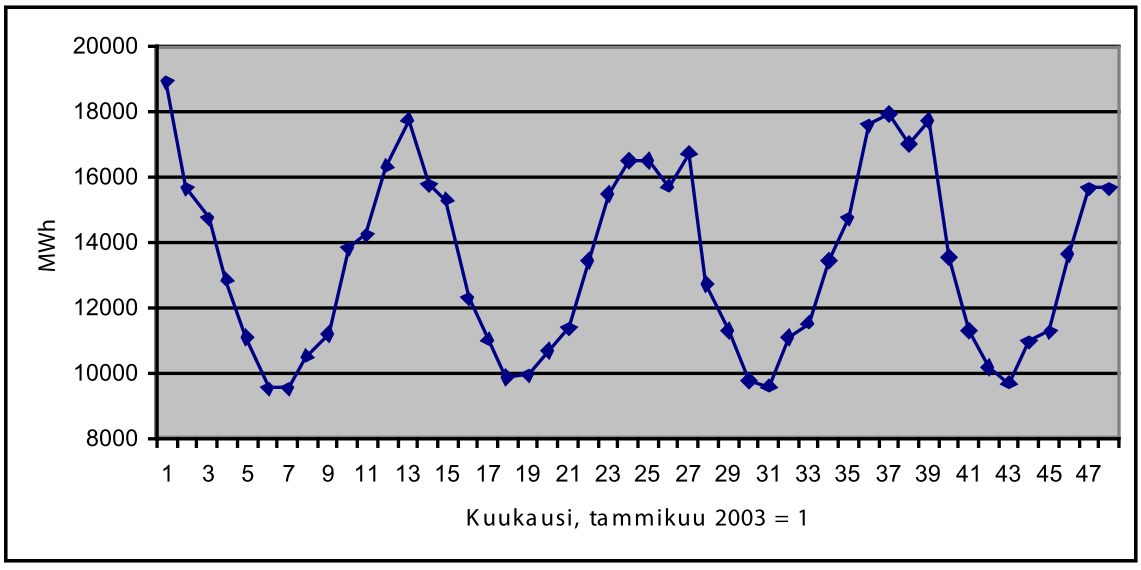
Aina ei ole helppoa havaita pelkkiä lukuja katselemalla, esiintyykö aikasarjassa kausivaihtelua vai ei. Alkeellisin vaihtoehto on graafinen menetelmä, jossa piirretään aikasarjan kuvaaja ja tutkitaan, onko aikasarjassa säännönmukaista vaihtelua siten, että säännönmukaisuus toistuu kiinteän jakson väliajoin.

ESIMERKKI 6

Valkeakosken Energia Oy:n sähköenergian siirto (MWh) oli vuosina 2003–2006 kuukausittain seuraava (lähde: Valkeakosken Energia Oy):

Siirto	Vuosi 2003	Vuosi 2004	Vuosi 2005	Vuosi 2006
Tammikuu	18 940	17 749	16 486	17 922
Helmikuu	15 647	15 758	15 690	17 020
Maaliskuu	14 762	15 296	16 736	17 704
Huhtikuu	12 879	12 263	12 721	13 533
Toukokuu	11 080	11 022	11 302	11 303
Kesäkuu	9 560	9 908	9 747	10 186
Heinäkuu	9 550	9 944	9 608	9 736
Elokuu	10 512	10 676	11 108	10 987
Syyskuu	11 187	11 349	11 543	11 264
Lokakuu	13 809	13 451	13 440	13 646
Marraskuu	14 208	15 487	14 717	15 681
Joulukuu	16 319	16 503	17 613	15 631
Yhteensä	158 453	159 406	160 711	164 613

Piirretään aluksi aikasarjan kuvaaja.



Valkeakosken Energia Oy:n sähköenergian siirto 2003–2006 (MWh). Alkuperäinen aikasarja.

Liukuvan keskiarvon käyttö

Yksinkertaista liukuvaa keskiarvoa (simple moving average) laskettaessa muodostetaan *k*:n peräkkäisen aikasarjan arvon summa, joka jaetaan *k*:lla. Tämän jälkeen jätetään ensimmäinen (vanhin) aikasarjan arvo pois. Loppuun lisätään seuraavaksi mukaan tuleva arvo ja lasketaan näiden lukujen keskiarvo. Näin jatketaan, kunnes kaikki aikasarjan arvot on käyty läpi. Saadut liukuvan keskiarvon arvot sijoitetaan kulloinkin laskemisessa käytettyjen arvojen ajalliseen keskikohtaan.

Liukuvan keskiarvon tarkoituksena on tasoittaa aikasarjaa. Keskeisenä kysymyksenä on tällöin vakion *k* arvo. Mitä pienempi arvo *k*:lla on, sitä vähäisempää tasoitus (smoothing) on ja sitä tarkemmin liukuvan keskiarvon arvot seuraavat alkuperäisiä aikasarjan arvoja (tracking). Vastaavasti mitä suuremmaksi luvuksi *k* valitaan, sitä enemmän tasoitusta saadaan aikaan mutta sitä vähemmän liukuvan keskiarvon arvot seuraavat alkuperäisen aikasarjan arvoja. Liukuvan keskiarvon käyttötarkoitus vaikuttaa valittavaan *k*:n arvoon.

Muunkinlaisia tasoitusmenetelmiä on kehitetty. Yksi suosituimmista on **eksponentiaalinen tasoitus** (exponential smoothing), jota

käytettäessä laskeminen jätetään tietokoneohjelman tehtäväksi mutta sopivien tasoitusparametrien löytäminen vaatii yleensä vuorovaikutteista kokeilua.

Kausivaihtelun vaikutusta aikasarjassa voidaan tarkastella käyttämällä liukuvaa keskiarvoa. Kun liukuvan keskiarvon laskemiseksi käytetään kausivaihtelun jaksoa L, saadaan aikasarjasta poistetuksi kausivaihtelun osuus, jolloin tuloksena on **kausipuhdistettu** aikasarja. Liukuva keskiarvo kuvaa tällöin vain trendiä ja suhdannevaihtelua. Jos liukuvaa keskiarvoa (moving average) merkitään MA:lla, saadaan

MA = T + C, joten Y – MA = S + I (summamalli)
MA = T · C, joten Y / MA = S · I (tulomalli).

Usein kausivaihtelua etsitään aikasarjoista, jotka on saatu tilastoimalla tulokset neljännesvuosittain tai kuukausittain. Jos liukuva keskiarvo lasketaan silloin, kun jaksoon sisältyy parillinen määrä arvoja, liukuvan keskiarvon tulos tulee kahden keskimmäisen arvon ajalliseen keskikohtaan. Kuvan tekemisen ja myös tulkinnan kannalta tämä on hieman hankalaa. Tällaisissa tapauksissa voidaankin menetellä siten, että aikasarjan arvoja painotetaan neljännesvuositilastoissa

$$\mathrm{MA}_t = \frac{0{,}5 \cdot y_{t-2} + y_{t-1} + y_t + y_{t+1} + 0{,}5 \cdot y_{t+2}}{4}$$

ja kuukausitilastoissa

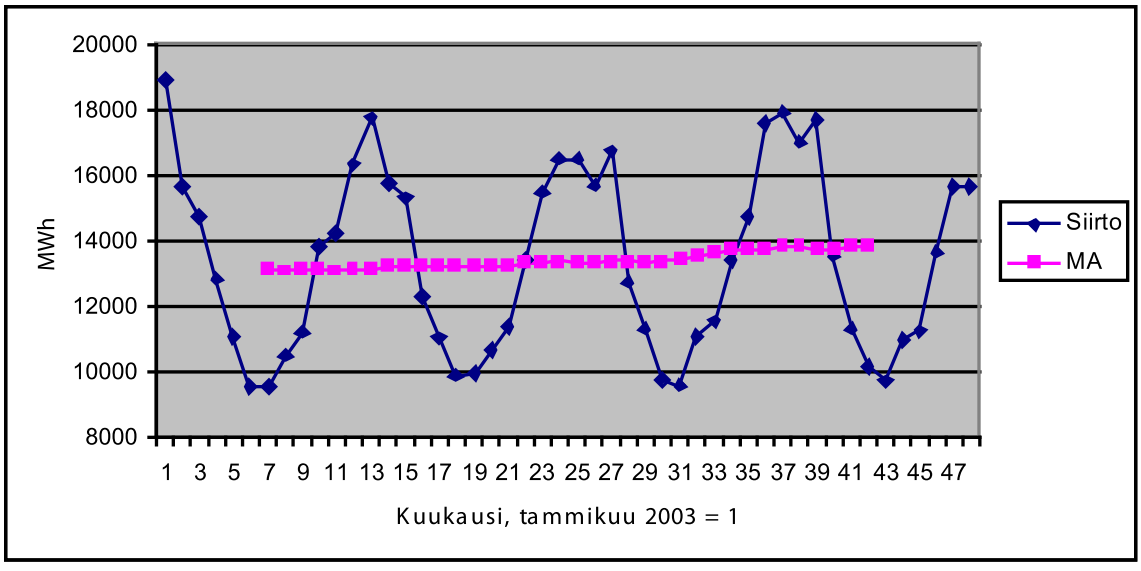
$$\mathrm{MA}_t = \frac{0{,}5 \cdot y_{t-6} + y_{t-5} + \ldots + y_{t+5} + 0{,}5 \cdot y_{t+6}}{12}.$$

Edellisessä tapauksessa otetaan viiden ja jälkimmäisessä tapauksessa kolmentoista luvun painotettu keskiarvo.

Toinen mahdollisuus on laskea ensin tavanomaiset liukuvat keskiarvot ja sen jälkeen vielä kahden perättäisen liukuvan keskiarvon keskiarvo. Tällöin esimerkiksi kuukausitilastoissa ensimmäinen luku tulee vastaamaan heinäkuun arvoa. Tätä menettelyä sanotaan **keskitetyksi liukuvaksi keskiarvoksi** (centered moving average).

ESIMERKKI 7

Määritetään esimerkin 6 aikasarjasta liukuvan keskiarvon kuvaaja, kun k = 12. Liukuvan keskiarvon arvot käyvät ilmi esimerkin 8 taulukosta.



Valkeakosken Energia Oy:n sähköenergian siirto 2003–2006 (MWh). Alkuperäinen aikasarja ja liukuvan keskiarvon kuvaaja.

Kausivaihteluindeksien laskeminen

Kausivaihtelua sisältävästä aikasarjasta haetaan aluksi kausivaihtelujakson pituus L. Kun jakson pituus on saatu selville, kuvataan jakson jokaista osaväliä kausivaihteluindeksin avulla.

Kausivaihteluindekseille S_i on voimassa

$$\sum_{i=1}^{L} S_i = 0 \qquad \text{(summamalli)}$$

$$\sum_{i=1}^{L} S_i = L \qquad \text{(tulomalli).}$$

Koska kausivaihtelun yhteydessä samat kausivaihteluindeksin arvot toistuvat aina jakson L välein, mallista riippumatta $S_i = S_{i+L} = S_{i+2L} = \ldots$ Tämä tarkoittaa sitä, että kuukausitilastoissa esimerkiksi tammikuun kausivaihteluindeksi on jokaisen vuoden osalta aina sama.

Ennen kuin kausivaihteluindeksiä ryhdytään laskemaan, aikasarjasta poistetaan aluksi mahdollinen trendin vaikutus vähentämällä arvoista MA tai jakamalla arvot MA:lla. Saadut tulokset järjestetään taulukoksi ja kunkin kuukauden tai neljännesvuoden indeksien keskiarvot lasketaan. Lisäksi pidetään huoli siitä, että summamallissa summa on nolla ja tulomallissa summa on jaksojen lukumäärä.

ESIMERKKI 8

Määritä esimerkin 6 aikasarjan kausivaihteluindeksit tulomallia varten.

Oheiseen taulukkoon on laskettu keskistetyt liukuvan keskiarvon arvot, joilla on jaettu alkuperäiset aikasarjan arvot. Saadut tulokset on merkitty taulukkoon, josta saadaan lopulliset kausivaihteluindeksit. Kausivaihteluindeksin käyttöön tutustutaan ennusteiden laatimisen yhteydessä esimerkissä 10.

Kausivaihteluindeksin laskeminen

t	Vuosi	Kuukausi	Siirto	Keskistetty MA	Suhdeluku
1	2003	Tammikuu	18 940		
2	2003	Helmikuu	15 647		
3	2003	Maaliskuu	14 762		
4	2003	Huhtikuu	12 879		
5	2003	Toukokuu	11 080		
6	2003	Kesäkuu	9 560		
7	2003	Heinäkuu	9 550	13 155	0,726
8	2003	Elokuu	10 512	13 110	0,802
9	2003	Syyskuu	11 187	13 137	0,852
10	2003	Lokakuu	13 809	13 133	1,051
11	2003	Marraskuu	14 208	13 105	1,084
12	2003	Joulukuu	16 319	13 117	1,244
13	2004	Tammikuu	17 749	13 148	1,350
14	2004	Helmikuu	15 758	13 171	1,196
15	2004	Maaliskuu	15 296	13 185	1,160
16	2004	Huhtikuu	12 263	13 177	0,931
17	2004	Toukokuu	11 022	13 215	0,834
18	2004	Kesäkuu	9 908	13 276	0,746
19	2004	Heinäkuu	9 944	13 231	0,752
20	2004	Elokuu	10 676	13 176	0,810

21	2004	Syyskuu	11 349	13 233	0,858
22	2004	Lokakuu	13 451	13 312	1,010
23	2004	Marraskuu	15 487	13 343	1,161
24	2004	Joulukuu	16 503	13 348	1,236
25	2005	Tammikuu	16 486	13 327	1,237
26	2005	Helmikuu	15 690	13 331	1,177
27	2005	Maaliskuu	16 736	13 357	1,253
28	2005	Huhtikuu	12 721	13 365	0,952
29	2005	Toukokuu	11 302	13 332	0,848
30	2005	Kesäkuu	9 747	13 346	0,730
31	2005	Heinäkuu	9 608	13 452	0,714
32	2005	Elokuu	11 108	13 568	0,819
33	2005	Syyskuu	11 543	13 663	0,845
34	2005	Lokakuu	13 440	13 737	0,978
35	2005	Marraskuu	14 717	13 771	1,069
36	2005	Joulukuu	17 613	13 790	1,277
37	2006	Tammikuu	17 922	13 813	1,297
38	2006	Helmikuu	17 020	13 814	1,232
39	2006	Maaliskuu	17 704	13 797	1,283
40	2006	Huhtikuu	13 533	13 794	0,981
41	2006	Toukokuu	11 303	13 843	0,817
42	2006	Kesäkuu	10 186	13 800	0,738
43	2006	Heinäkuu	9 736		
44	2006	Elokuu	10 987		
45	2006	Syyskuu	11 264		
46	2006	Lokakuu	13 646		
47	2006	Marraskuu	15 681		
48	2006	Joulukuu	15 631		

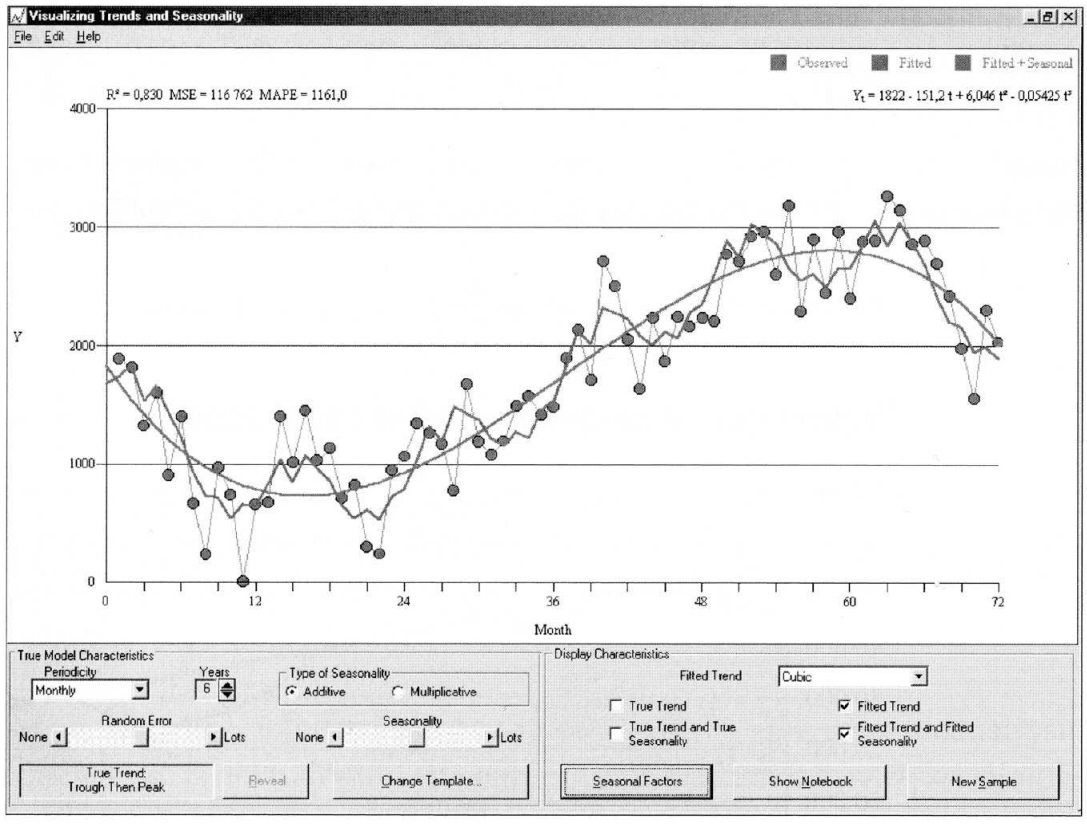
Lopulliset kausivaihteluindeksit

	1	2	3	4	5	6
2003						
2004	1,350	1,196	1,160	0,931	0,834	0,746
2005	1,237	1,177	1,253	0,952	0,848	0,730
2006	1,297	1,232	1,283	0,981	0,817	0,738
Keskiarvo	1,295	1,202	1,232	0,955	0,833	0,738
Sovitettu ka.	**1,293**	**1,200**	**1,230**	**0,953**	**0,832**	**0,737**

	7	8	9	10	11	12
2003	0,726	0,802	0,852	1,051	1,084	1,244
2004	0,752	0,810	0,858	1,010	1,161	1,236
2005	0,714	0,819	0,845	0,978	1,069	1,277
2006						
Keskiarvo	0,731	0,810	0,852	1,013	1,105	1,252
Sovitettu ka.	**0,730**	**0,809**	**0,850**	**1,012**	**1,103**	**1,251**

10.4 Suhdannevaihtelu

Suhdannevaihtelu on pitkän aikavälin heilahtelua, ja sitä kuvaa aaltomainen käyrä, jonka aallonpituus on useita vuosia. Aaltomainen heilahtelu ei ole tavallisesti kovin säännöllistä, minkä vuoksi suhdannevaihtelukomponenttia on vaikea mallintaa. Jos aikasarjassa esiintyy suhdannevaihtelua, siinä havaitaan eri vaiheita, esimerkiksi noususuhdanne, korkeasuhdanne, laskusuhdanne ja matalasuhdanne.



Aikasarja, kausivaihtelu ja suhdannevaihtelu Visual Statistics -ohjelmalla.

Suhdannevaihtelu voidaan parhaiten tunnistaa aikasarjasta, jos siitä ensin poistetaan trendin ja kausivaihtelun vaikutukset:

$$Y - T - S = C + I \quad \text{(summamalli)}$$

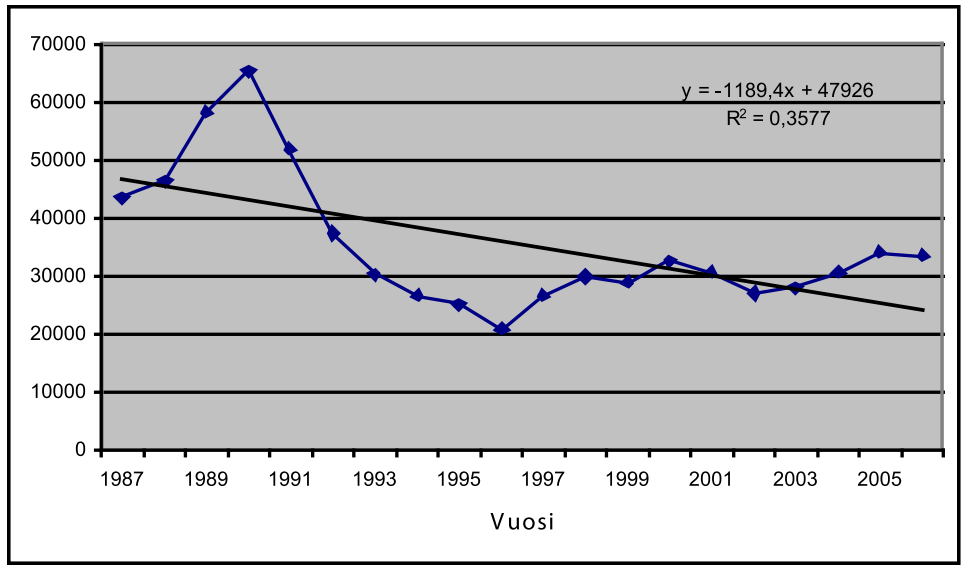
$$\frac{Y}{T \cdot S} = C \cdot I \quad \text{(tulomalli).}$$

Jäljelle jäänyt aikasarja kuvataan graafisesti ja pyritään löytämään siitä säännönmukaisuutta. Jos kuvassa on säännönmukaisuutta, se tulkitaan osoitukseksi siitä, että aikasarjassa on suhdannevaihtelua mukana. Aina aikasarjassa ei kuitenkaan ole selkeästi trendiä tai kausivaihtelua, jolloin on tyydyttävä tutkimaan pelkästään alkuperäistä aikasarjaa tai siitä muunnoksien avulla saatuja kuvaajia.

Usein jää huomaamatta, että jos aikasarja esitetään vuoden mittaisena tilastona, aikasarjassa ei voi olla lainkaan suhdannevaihtelua, koska suhdannevaihtelun jakson pituus on aina yli vuosi. Toisaalta jos aikasarja esittää vain vuosittaisia tuloksia, aikasarjassa ei voi olla kausivaihtelua, sillä kausivaihtelun jakson pituus on korkeintaan yksi vuosi.

ESIMERKKI 9

Rakennusteollisuutta pidetään suhdanneherkkänä. Tutkitaan kuvan avulla, havaitaanko vuosina 1987–2006 valmistuneiden asuinhuoneistojen määrässä suhdannevaihtelua (lähde: Tilastokeskus). Piirretään alkuperäisen aikasarjan kuvaaja ja siihen trendi.



Valmistuneet asuinhuoneistot 1987–2006.

Tulkinta: Kuvaajassa voidaan havaita tiettyä säännönmukaisuutta, joka voitaneen selittää suhdannevaihteluksi. Sen mallintaminen ei kuitenkaan ole mahdollista käytettävissä olevilla tiedoilla.

Suhdannevaihtelun aiheuttamat muutokset aikasarjaan pystytään toisinaan arvioimaan kiertoteitse. Pyrkimyksenä on tällöin löytää toinen aikasarja, jonka suhdannevaihtelut noudattavat suurin piirtein alkuperäisen aikasarjan suhdannevaihtelua. Jos aikasarjoissa esiintyy viive siten, että ensimmäisessä aikasarjassa havaittu heilahtelu toistuu jonkin asteisena toisessa aikasarjassa tietyn ajan kuluttua, tätä havaintoa voidaan hyödyntää. On esimerkiksi havaittu, että Yhdysvalloissa tapahtuvat suhdannevaihtelut heijastuvat usein Suomeenkin jonkin ajan kuluttua.

10.5 Satunnaisvaihtelu ja jäännössarja

Satunnaisvaihtelu on nimensä mukaisesti satunnaista. Joissakin aikasarjoissa satunnaisvaihtelun osuus on suuri, toisissa vähäinen. Satunnaisvaihtelun suuruutta voidaan kuitenkin jossakin määrin arvioida jäännössarjan perusteella. Kun aikasarjasta poistetaan trendi, kausivaihtelu ja suhdannevaihtelu, jäljelle jää pelkästään satunnaisvaihtelua. Jos näin saatavassa jäännössarjassa esiintyy pelkästään satunnaista heilahtelua, muut komponentit sisältävät kaiken aikasarjasta saatavissa olevan tiedon. Muussa tapauksessa komponentteja ei ole saatu määritetyksi riittävän tarkasti.

10.6 Ennusteiden laatiminen

Yksinkertaisin tapa laatia ennuste on tehdä graafinen esitys, jossa aikasarjan kuvaajaa jatketaan silmämääräisesti tunnettujen tosiasioiden ja oman näkemyksen mukaisesti, minkä jälkeen luetaan saadusta kuvaajasta ennustearvot.

Matemaattisten ennusteiden laatimiseen tutustuttiin regressioanalyysin yhteydessä. Tuolloin todettiin, että ennustearvo saadaan määritetyksi sijoittamalla malliin selittävien muuttujien paikalle ennusteessa käytettävät arvot ja laskemalla näillä arvoilla selitettävän muuttujan y arvo. Samalla tavalla menetellään aikasarjojen yhteydessä. Aikasarjojen mallit voivat kuitenkin muodostua aikaisempia monimutkaisemmik-

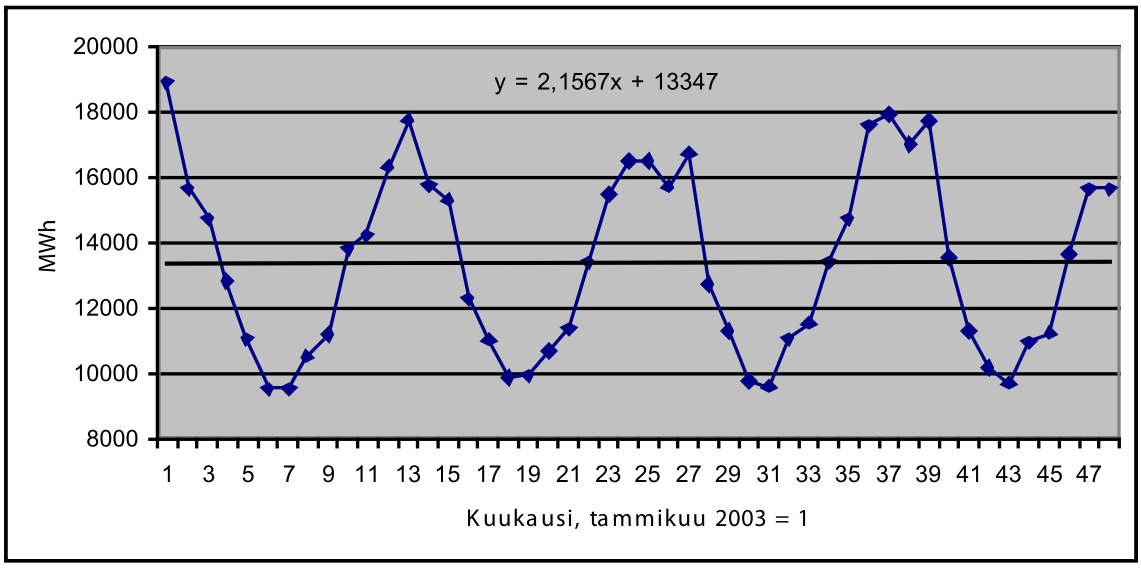
sikin, koska malli muodostuu komponenteista ja kukin komponentti on otettava erikseen huomioon.

Seuraavassa osoitetaan, kuinka ennuste laaditaan silloin, kun käytetään klassisen aikasarja-analyysin mukaista mallia.

ESIMERKKI 10

Laaditaan esimerkin 6 tilastoaineistosta tulomallia käyttämällä vuoden 2007 ennuste.

Aluksi tarvitaan suoraviivaisen trendin yhtälö.



Valkeakosken Energia Oy:n sähköenergian siirto 2003–2006 (MWh). Aikasarjan kuvaaja, trendisuora ja sen yhtälö.

Kuvasta käy ilmi, että trendisuoran yhtälö on

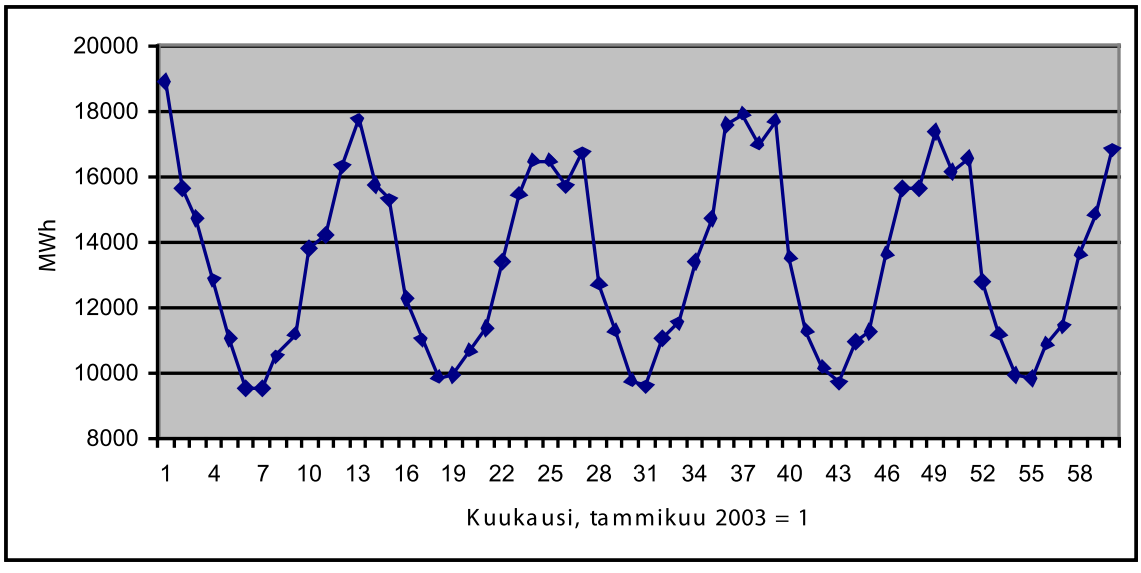
$y = 13\,347 + 2{,}1567 \cdot t$ ($t = 1$ tammikuussa 2003).

Kun verrataan esimerkin 7 kuvaa sivulla 322 ja yllä olevaa kuvaa, voidaan havaita, että trendi pitää sisällään melko tarkkaan myös suhdannevaihtelukomponentin. Koska saatavilla ei ole keinoja satunnaisvaihtelun huomioon ottamiseksi, tulomalliksi saadaan

$Y_t = T_t \cdot S_t$.

Kun kausivaihteluindeksit poimitaan esimerkistä 8, vuoden 2007 ennustearvoiksi saadaan seuraavat tulokset:

Tammikuu	(13 347 + 2,1567 · 49) · 1,293 =	17 394
Helmikuu	(13 347 + 2,1567 · 50) · 1,200 =	16 146
Maaliskuu	(13 347 + 2,1567 · 51) · 1,230 =	16 552
Huhtikuu	(13 347 + 2,1567 · 52) · 0,953 =	12 827
Toukokuu	(13 347 + 2,1567 · 53) · 0,832 =	11 200
Kesäkuu	(13 347 + 2,1567 · 54) · 0,737 =	9 923
Heinäkuu	(13 347 + 2,1567 · 55) · 0,730 =	9 830
Elokuu	(13 347 + 2,1567 · 56) · 0,809 =	10 895
Syyskuu	(13 347 + 2,1567 · 57) · 0,850 =	11 449
Lokakuu	(13 347 + 2,1567 · 58) · 1,012 =	13 634
Marraskuu	(13 347 + 2,1567 · 59) · 1,103 =	14 862
Joulukuu	(13 347 + 2,1567 · 60) · 1,251 =	16 859.



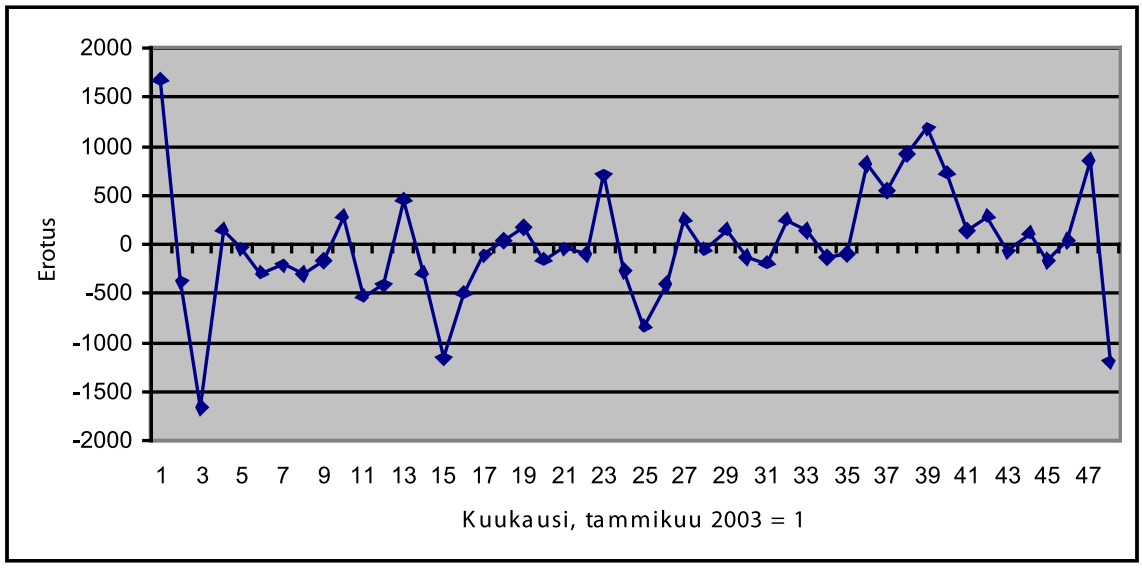
Valkeakosken Energia Oy:n sähköenergian siirto 2003–2006 ja ennuste vuodelle 2007. Alkuperäistä aikasarjaa on jatkettu ennusteen mukaisesti.

Tulkinta: Kun verrataan tuloksia vuoden 2006 arvoihin, voidaan havaita, että ennustearvot ovat jotakuinkin samansuuruisia kuin vuonna 2006, joskin pientä kasvua on ennakoitavissa. Tämä selittyy siitä, että tilastoaineistossa on vain hyvin loiva nouseva trendi. Toisaalta mikään ennustearvo ei ole suhteellisesti eikä absoluuttisestikaan kovin kaukana vuoden 2006 vastaavan ajankohdan arvosta, joten tuloksia voidaan pitää sekä mielekkäinä että mahdollisina. Ennustetta laadittaessa ei myöskään ole ollut tiedossa mitään muita seikkoja, jotka voisivat vaikuttaa ennusteisiin. Näitä voisivat olla esimerkiksi kunnan aluemuutokset, teollisuuslaitosten laajennukset, lisäykset tai vähentymiset, ulkovalaistuksen voimakas lisääminen tai vähentäminen ja sähkölämmitystalouksien määrän selkeä lisääntyminen tai väheneminen. Sähköenergian siirtoa kuvaava käyrä noudattaa ennusteessa varsin hyvin aikaisempien vuosien kehitystä.

Ennusteen laatiminen tietokoneohjelmalla

Monien tietokoneohjelmien avulla voidaan laatia lähes automaattisesti aikasarjojen ennusteita. Asiaa varten on kehitetty useita eri menetelmiä, joiden antamat ennusteet poikkeavat hieman toisistaan. Vaikka tehokkaiden tietokoneohjelmien antamia ennusteita pidetään klassista komponenttimallia kehittyneempänä, klassinen komponenttimalli ei ole suinkaan huono suhteellisen säännöllisten aikasarjojen ennusteiden laatimisessa.

Ennusteiden luotettavuus

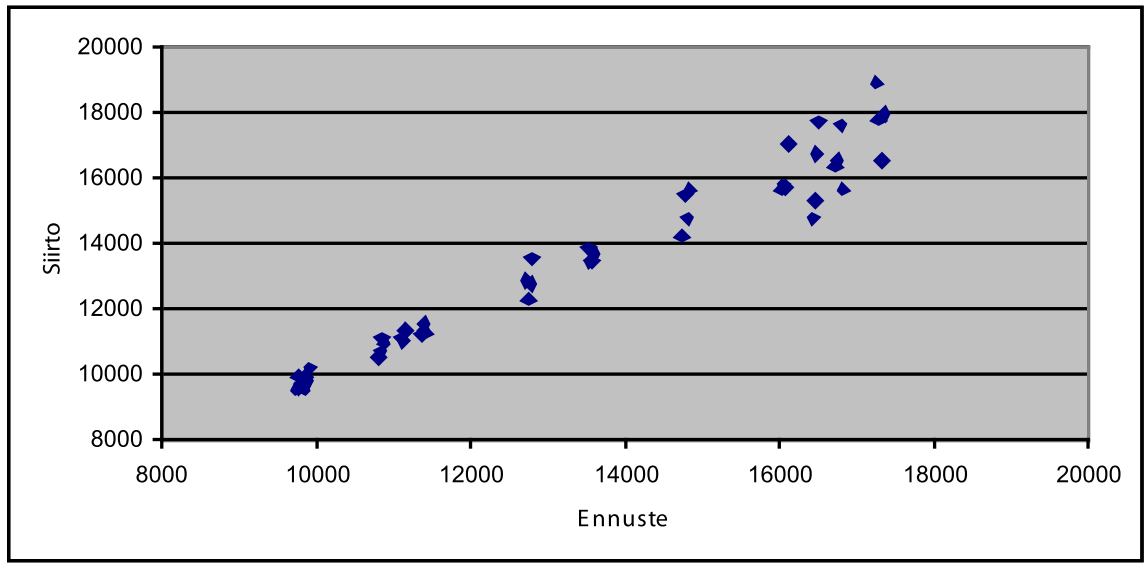


Alkuperäisen aikasarjan arvojen ja esimerkissä 10 laaditun mallin ennustearvojen erotusten kuvaaja.

Jos malli selittää hyvin alkuperäisen aikasarjan käyttäytymistä, saadaan yleensä varsin luotettavia ennusteita. Mallin hyvyyttä tutkitaan samaan tapaan kuin regressioanalyysissä. Vähimmäismenettelynä pidetään sitä, että piirretään kuvaaja pisteiden (t, e_t) kautta, kun $e_t = y_t - \hat{y}_t$. Jos malli on oikein määritelty, kuvaajassa ei pitäisi näkyä mitään säännönmukaisuutta. Jos säännönmukaisuutta kuitenkin esiintyy, mallia on yritettävä parantaa.

Tulkinta: Todellisten arvojen ja ennustearvojen erotuksen kuvaaja heilahtelee satunnaisesti nollan molemmin puolin, eikä kuvaajassa esiinny selkeää säännönmukaisuutta. Näyttää siis siltä, ettei mallissa ole systemaattista virhettä. Myös erotukset e_t ovat varsin pieniä, joten käytetty malli kuvaa aikasarjaa melko hyvin.

Ennusteiden luotettavuutta voidaan toisaalta tutkia graafisesti samalla tavalla kuin aikaisemminkin: Merkitään pisteet $(\hat{y}_t, y_t)$ koordinaatistoon ja tarkastellaan kuvaajaa. Malli antaa todennäköisesti sitä tarkempia ennusteita, mitä tarkemmin pisteet muodostavat suoran viivan.



Esimerkissä 10 laaditun mallin antaman ennusteen hyvyyden graafinen tarkastelu.

Tulkinta: Pisteet $(\hat{y}_t, y_t)$ sijaitsevat lähellä yllä olevan kuvan kuviteltua lävistäjää, mikä osoittaa, että malli sopii havaintoarvoihin melko hyvin. Tällöin voidaan olettaa, että myös ennustearvot olisivat varsin tarkkoja. Tämä ei kuitenkaan tarkoita sitä, ettei jokin toinen malli voisi olla vieläkin parempi ja ettei sen avulla saataisi vieläkin tarkempia ennustearvoja.

10.7 Case D: Vanhuuseläkeläisten määrän kehittyminen

Tutkimuksen tausta

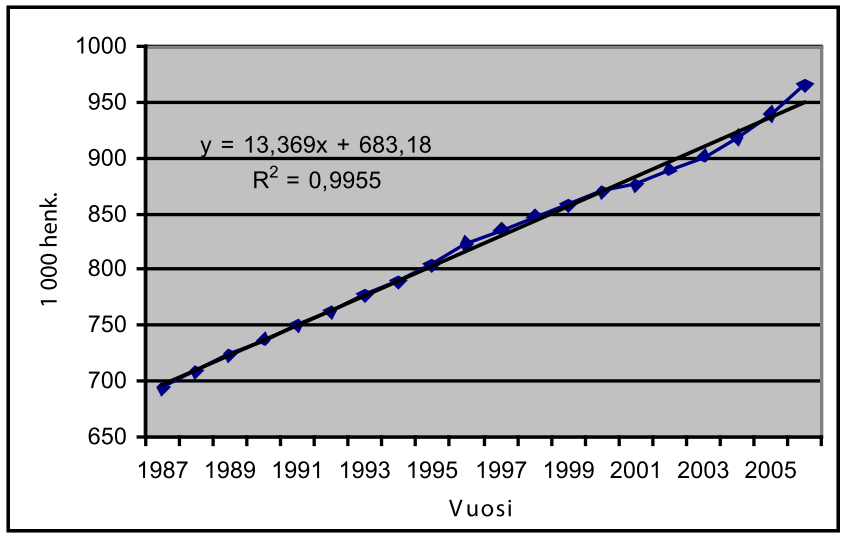
Tutkittiin, miten vanhuuseläkeläisten määrä on kehittynyt Suomessa vuodesta 1987 vuoteen 2006 mennessä ja kasvaako vanhuuseläkeläisten määrä selvästi. Samalla haluttiin laatia ennusteet kahdesta seuraavasta vuodesta ja vuodesta 2010. Selvitystä varten saatiin tilastoaineisto Tilastokeskuksesta. Seuraavassa taulukossa vanhuuseläkeläisten määrä esitetään tuhansina. Selvitystyössä tarvittavat sarakkeet on lisätty valmiiksi taulukkoon.

Vanhuuseläkeläisten määrä Suomessa 1987–2006

Vuosi	Vanhuksia, 1 000 henk.	Suhdeluku	Erotus	Ennuste	*e*
1987	694,2			696,5	–2,3
1988	709,1	1,021	14,9	709,9	–0,8
1989	723,9	1,021	14,8	723,3	0,6
1990	737,2	1,018	13,3	736,7	0,5
1991	750,9	1,019	13,7	750,0	0,9
1992	762,6	1,016	11,7	763,4	–0,8
1993	776,8	1,019	14,2	776,8	0,0
1994	789,4	1,016	12,6	790,1	–0,7
1995	804,1	1,019	14,7	803,5	0,6
1996	822,5	1,023	18,4	816,9	5,6
1997	836,7	1,017	14,2	830,2	6,5
1998	846,9	1,012	10,2	843,6	3,3
1999	858,2	1,013	11,3	857,0	1,2
2000	869,7	1,013	11,5	870,3	–0,6
2001	875,6	1,007	5,9	883,7	–8,1
2002	889,6	1,016	14,0	897,1	–7,5
2003	901,9	1,014	12,3	910,5	–8,6
2004	917,3	1,017	15,4	923,8	–6,5
2005	940,0	1,025	22,7	937,2	2,8
2006	964,4	1,026	24,4	950,6	13,8

Kuvaaja

Laaditaan aluksi aikasarjan kuvaaja ja piirretään siihen suoraviivainen trendi.



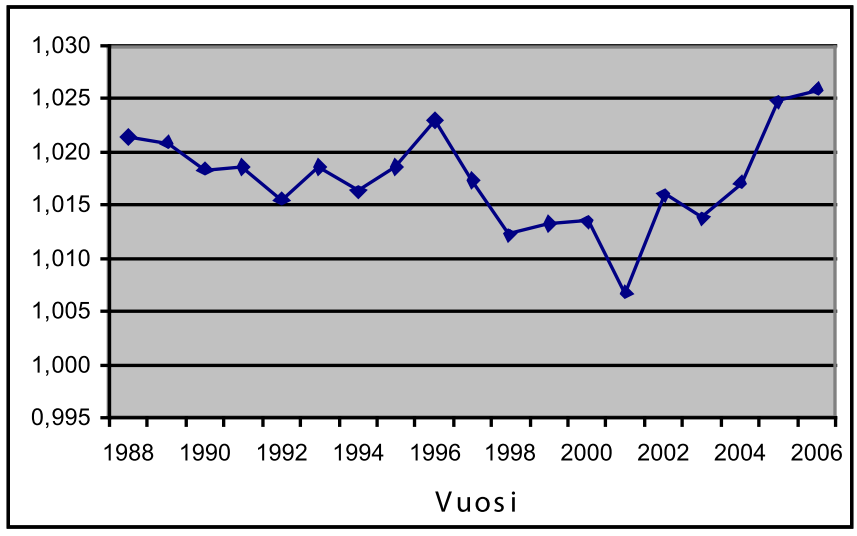
Vanhuuseläkeläisten määrä Suomessa 1987–2006. Vanhuuseläkeläisten määrän kuvaaja trendisuorineen.

Kuvaajasta käy ilmi, että vanhuuseläkeläisten määrä on kasvanut varsin tasaisesti vuodesta 1987 vuoteen 2006.

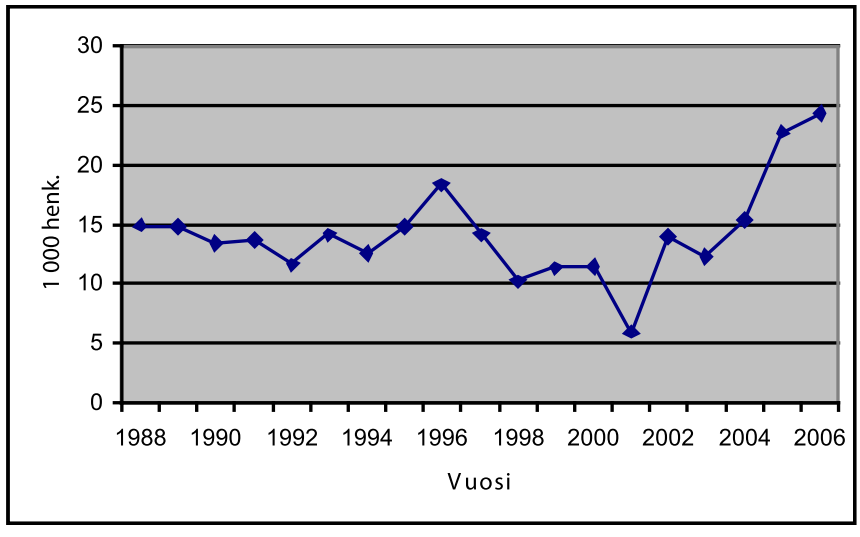
Muunnokset

Seuraavaksi harkitaan, voidaanko aikasarjaa jollakin muunnoksella muokata sopivampaan muotoon. Aluksi todetaan, ettei indeksien käytöstä sinänsä ole hyötyä, koska kyseessä ei ole vertailu mihinkään kiinteään ajankohtaan eikä myöskään usean aikasarjan vertailu. Toisaalta koska aikasarjan kuvaaja ei kaareudu selvästi ylös- eikä alaspäin, logaritmi- tai neliöjuurimuunnoskaan ei auta. Peräkkäisten termien suhde ja erotus kannattaa sen sijaan laskea. Tulokset esitettiin edellä olevassa taulukossa Suhdeluku- ja Erotus-sarakkeissa.

Taulukosta käy ilmi, että peräkkäisten termien suhde osoittaa aluksi runsaan kahden prosentin kasvua vuodessa ja lopussa vuotuinen kasvu on noin 2,5 prosenttia. Peräkkäisten termien erotus pysyy koko ajan positiivisena ja vaihtelee viiden ja kahdenkymmenenviiden tuhannen välillä, mikä puoltaa vahvasti nousevan trendin olemassaoloa.



Vanhuuseläkeläisten määrän kehittyminen suhdelukuna. Peräkkäisten lukumäärien suhteiden kuvaaja.



Vanhuuseläkeläisten määrän kehittyminen erotuksena edellisestä vuodesta. Peräkkäisten lukumäärien erotuksen kuvaaja.

Koska suora viiva näyttää sivun 333 kuvan perusteella mahdolliselta mallilta eikä muita vaihtoehtoja ole tullut esille, seuraavaksi tehdään lineaarinen regressioanalyysi.

Sivun 333 kuvaajasta käy ilmi, että regressiosuoran yhtälö on

$y = 683{,}18 + 13{,}369 \cdot t$ (t = 1 vuonna 1987)

tai

$y = 683{,}18 + 13{,}369 \cdot (t - 1986)$ (t = tarkasteltava vuosiluku).

Seuraavaksi selvitetään, onko vanhuuseläkeläisten määrä kiistatta kasvussa. Tämä saadaan selville testaamalla, onko regressiosuoran kulmakerroin positiivinen.

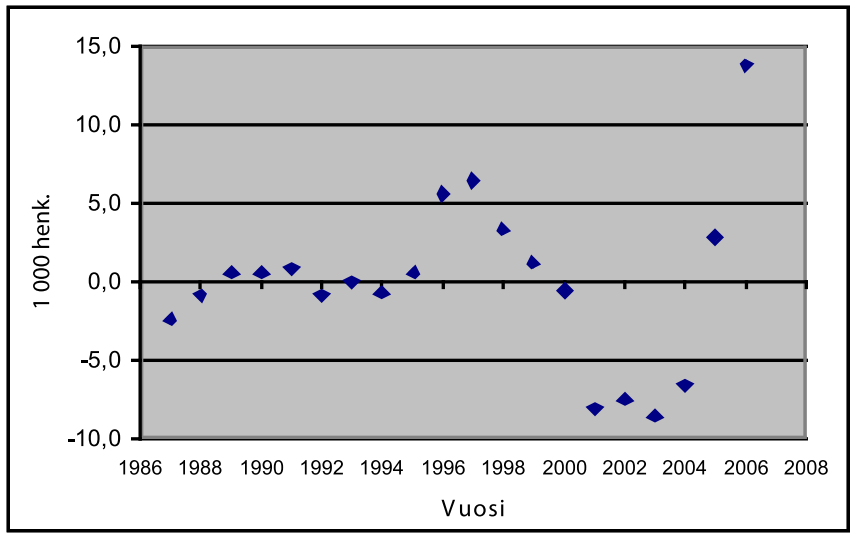
		Unstandardized Coefficients		Standardized Coefficients		
Model		**B**	**Std. Error**	**Beta**	***t***	**Sig.**
1	(Constant)	683,175	2,526		270,434	,000
	Kausi	13,369	,211	,998	63,395	,000

Regressioanalyysin tuloste.

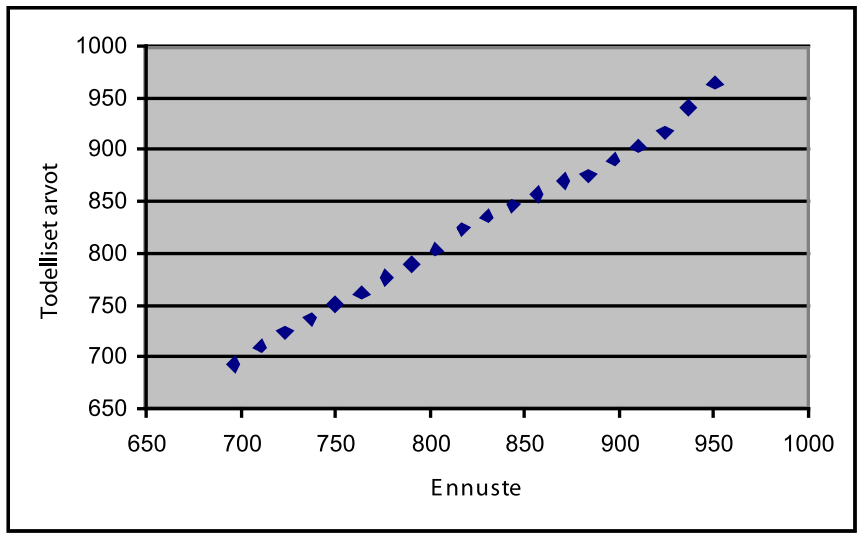
Yllä olevasta taulukosta käy ilmi, että kulmakertoimen *t*-arvo on 63,395 > 2, mikä osoittaa ilman *t*-jakauman taulukkoakin, että kulmakerroin poikkeaa nollasta erittäin merkitsevästi. Vanhuuseläkeläisten määrä on siis kasvussa.

Mallin hyvyys

Seuraavaksi tutkitaan mallin hyvyyttä. Asian selvittämiseksi lasketaan aluksi regressiosuoran antamat ennustearvot vuosista 1987–2006. Nämä arvot on merkitty sivun 332 taulukon Ennuste-sarakkeeseen. Tulosten perusteella lasketaan jäännöstermit eli $e_t = y_t - \hat{y}_t$, jolloin mallin hyvyyttä voidaan tarkastella kuvaajien avulla.



Todellisten arvojen ja ennustearvojen erotusten kuvaaja. Mallin hyvyyden tutkiminen graafisesti.



Ennustearvojen ja todellisten arvojen tarkastelua. Mallin hyvyyden tutkiminen graafisesti.

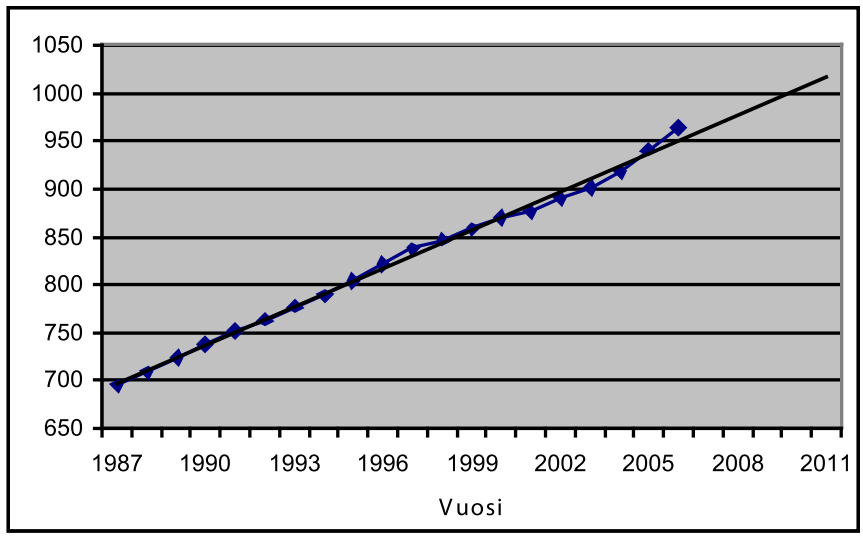
Sivun 335 kuvaajasta käy ilmi, että jäännöstermeissä on jonkin verran säännönmukaisuutta. Tämä osoittaa, ettei suora viiva kuvaa täysin moitteettomasti aikasarjaa, joten ennusteissa saattaa esiintyä systemaattista virhettä eli **harhaa**. Toisaalta jälkimmäinen kuvaaja osoittaa, että mallin avulla lasketut arvot ja todelliset arvot ovat varsin lähellä toisiaan. Tämä voidaan havaita siitä, että jälkimmäisen kuvion pisteet ovat lähellä kuvion lävistäjää (vaikkei sitä olekaan piirretty näkyviin). Voidaan siis uskoa, että ennustearvot ovat kuitenkin todennäköisesti suhteellisen tarkkoja. Sama käsitys saadaan myös selityskertoimesta R^2, joka on niinkin suuri kuin 99,6 %.

Ennusteet

Seuraavaksi laaditaan ennusteet käyttämällä mallina regressiosuoraa.

Vuosi 2007: 683,18 + 13,369 · (2007–1986) = 964
Vuosi 2008: 683,18 + 13,369 · (2008–1986) = 977
Vuosi 2010: 683,18 + 13,369 · (2010–1986) = 1 004

Koska aikasarjan luvut on esitetty tuhansina, kyseisten vuosien pisteestimaatit ovat 964 000, 977 000 ja 1 004 000 henkilöä. Seuraavan sivun kuvasta voi tarkistaa graafisesti vuosien 2007, 2008 ja 2010 ennusteet.



Ennustearvojen graafinen toteaminen.

Tulkinta

Näyttää siltä, että lineaarinen malli kuvaa aikasarjaa varsin hyvin. Kasvu on ollut edellisinä vuosina vakaata, eikä suuria muutoksia ole odotettavissa. Trendisuoran kulmakertoimen testaus osoitti, että vanhuuseläkeläisten määrä todella on kasvussa.

Ennusteiden tarkkuutta puoltaa mallin korkea selityskerroin ja pienet jäännöstermit. Parin viimeisen vuoden aikana jäännöstermit ovat kuitenkin olleet positiivisia, mikä antaa aiheen olettaa, että vuosien 2007 ja 2008 ennusteet olisivat pikemminkin hieman liian pieniä kuin liian suuria.

Ennusteen laatiminen kauas tulevaisuuteen ei onnistu pelkän lineaarisen regression avulla, sillä liian monet seikat voivat huomattavasti vaikuttaa vanhuuseläkkeen saajien lukumäärään. Tällaisia seikkoja ovat esimerkiksi terveyden- ja sairaanhoidon kehittyminen, elämäntapojen ja eläkeiän mahdollinen muuttuminen sekä muuttoliikkeet. Lisäksi vanhuuseläkeläisten määrään vaikuttaa ikäluokkien koko, joka ei suinkaan ole vakio. Pelkästään edellä olevien tietojen perusteella ei siten ole lainkaan mielekästä esittää lineaarista regressiomallia käyttämällä kovin pitkälle tulevaisuuteen ulottuvaa ennustetta.

Helposti jää huomaamatta, että vaikka vanhuuseläkeläisten määrä on kasvanut, myös koko väestömäärä on kasvanut. Niinpä on aiheellista laskea myös vanhuuseläkeläisten suhteellinen vuosittainen osuus väestöstä.

Vanhuuseläkkeellä olevan väestön suhteellinen osuus ja kasvu

Vuosi	Vanhuksia, 1000 henk.	Koko väestö, 1 000 henk.	%	Suhdeluku
1987	694,2	4 939	14,1	
1988	709,1	4 954	14,3	1,021
1989	723,9	4 974	14,6	1,021
1990	737,2	4 989	14,8	1,018
1991	750,9	5 029	14,9	1,019
1992	762,6	5 055	15,1	1,016
1993	776,8	5 078	15,3	1,019
1994	789,4	5 099	15,5	1,016
1995	804,1	5 117	15,7	1,019
1996	822,5	5 132	16,0	1,023
1997	836,7	5 147	16,3	1,017
1998	846,9	5 160	16,4	1,012
1999	858,2	5 171	16,6	1,013
2000	869,7	5 181	16,8	1,013
2001	875,6	5 195	16,9	1,007
2002	889,6	5 206	17,1	1,016
2003	901,9	5 220	17,3	1,014
2004	917,3	5 237	17,5	1,017
2005	940,0	5 256	17,9	1,025
2006	964,4	5 277	18,3	1,026

Taulukosta käy ilmi, että vanhuuseläkkeellä olevia on nyt noin kuudesosa koko väestömäärästä. Vanhuuseläkkeellä olevien suhteellinen osuus on kasvanut vuosittain 0,7–2,6 % ja viime vuosina noin 2,5 %, eli kokonaiskasvu on ollut 20 vuodessa noin 4 prosenttiyksikköä.

TEHTÄVIÄ

1. Pohdi, mihin tilanteisiin esimerkin 1 eri kuvaajat soveltuvat parhaiten.

2. Seuraava taulukko esittää Suomen energian kokonaiskulutuksen terajouleina energialähteittäin vuosilta 1995–2006 (lähde: Tilastokeskus).

Vuosi	Öljy	Hiili	Maakaasu	Ydinenergia
1995	347 060	167 107	117 648	197 760
1996	356 335	205 532	123 084	203 771
1997	353 193	190 028	121 104	218 738
1998	364 601	147 125	138 744	228 829
1999	366 680	148 871	138 888	240 655
2000	353 607	148 375	141 876	235 364
2001	360 070	167 083	153 936	238 407
2002	365 535	183 589	152 856	233 400
2003	373 915	243 311	169 200	238 145
2004	373 819	218 966	163 008	237 970
2005	362 161	129 224	149 148	243 887
2006	361 079	217 706	160 560	239 804

Piirrä aikasarjojen kuvaajat.

3. Seuraava taulukko esittää pääkaupunkiseudun vanhojen kerrostalohuoneistojen keskimääräiset velattomat neliöhinnat vuosilta 1990–2006 (lähde: Tilastokeskus).

		Nimellishintaindeksi	Reaalihintaindeksi
Vuosi	**€/m²**	**1983 = 100**	
1990	1 704	209,5	145,1
1991	1 399	169,8	112,9
1992	1 082	129,9	84,2
1993	1 029	125,3	79,6
1994	1 142	138,9	87,2
1995	1 078	128,3	79,8
1996	1 124	137,8	85,2
1997	1 396	170,5	104,2
1998	1 545	189,9	114,4
1999	1 727	212,6	126,6
2000	1 916	232,3	133,8
2001	1 907	229,6	128,9
2002	2 112	254,3	140,7
2003	2 247	269,4	147,6
2004	2 386	285,7	156,3
2005	2 535	302,5	164,1
2006	2 783	332,1	177,1

Piirrä nimellishintaindeksin ja reaalihintaindeksin kuvaajat. Kokeile erilaisia muunnoksia. Mitä johtopäätöksiä voidaan tehdä?

4. Määritä tehtävän 2 tilastoaineistosta suoraviivaisen trendin yhtälö öljylle, hiilelle, maakaasulle ja ydinenergialle.

5. Määritä tehtävän 3 tilastoaineistosta suoraviivaisen trendin yhtälö vuosien 1996–2006 keskimääräisten velattomien neliöhintojen reaaliselle kasvulle.

6. Testaa, onko tehtävän 2 ydinenergian kulutusta esittävässä aikasarjassa nouseva lineaarinen trendi, kun merkitsevyystasoksi valitaan 5 %.

7. Testaa, onko tehtävän 3 tilastoaineistossa suoraviivainen nouseva trendi vuosien 1996–2006 pääkaupunkiseudun vanhojen kerrostalohuoneistojen keskimääräisten velattomien neliöhintojen nimelliselle kasvulle, kun merkitsevyystasoksi valitaan 5 %.

8. Testaa, onko esimerkin 6 aikasarjassa nouseva lineaarinen trendi, kun merkitsevyystasoksi valitaan 5 %.

9. Energia Oy:n sähköenergian hankinta (MWh) oli vuosina 2003–2006 kuukausittain seuraava:

Kuukausi/Vuosi	2003	2004	2005	2006
Tammikuu	20 758	22 203	23 859	25 793
Helmikuu	19 860	20 894	21 925	23 116
Maaliskuu	20 710	19 882	22 126	25 439
Huhtikuu	16 421	15 704	17 851	19 542
Toukokuu	13 550	14 387	16 592	17 388
Kesäkuu	13 296	11 686	14 886	15 030
Heinäkuu	11 579	11 360	14 419	15 163
Elokuu	12 936	12 981	16 276	17 104
Syyskuu	14 122	13 809	17 436	18 448
Lokakuu	17 202	17 078	19 299	21 765
Marraskuu	20 353	20 928	27 223	28 118
Joulukuu	21 530	23 887	29 688	33 234

Piirrä aikasarjan kuvaaja. Määritä lineaarisen trendin yhtälö ja testaa, onko trendin kulmakerroin nolla, kun merkitsevyystasoksi valitaan 1 %.

10. Määritä tehtävän 9 aikasarjasta kausivaihteluindeksit tulomallia varten.

11. Oheinen taulukko esittää autokaupan määräindeksit (2000 = 100) Suomessa vuosilta 1996–2006 (lähde: Tilastokeskus).

Vuosi	Autokauppa	Vuosi	Autokauppa
1996	73,6	2001	96,8
1997	81,9	2002	103,2
1998	93,2	2003	120,1
1999	99,5	2004	129,6
2000	100,0	2005	136,6
		2006	147,0

Määritä lineaarisen trendin yhtälö ja laadi sen perusteella vuoden 2010 ennuste. Mieti ennusteen mielekkyyttä.

12. Kuolemanvaaraluku ilmoittaa, kuinka monta promillea kaikista vuoden alussa elossa olevista samana vuonna syntyneistä kuolee todennäköisesti vuoden aikana. Suomessa miesten kuolemanvaaraluvut ikäryhmittäin ovat seuraavat (lähde: Tilastokeskus):

Ikä	Kuolemanvaara	Ikä	Kuolemanvaara
20	1,2	50	11,1
25	1,5	55	18,0
30	2,0	60	26,2
35	2,9	65	41,8
40	4,4	70	62,7
45	7,9	75	88,2

Kuvaa kuolemanvaaraa graafisesti sopivalla tavalla. Laadi 80-vuotiaita miehiä koskeva lineaarinen ennuste. Vertaa saamaasi tulosta todelliseen lukuun, joka on 138,1. Onko tilastoaineistoon olemassa jokin muu sopiva matemaattinen malli kuin suora viiva?

13. Oheisessa taulukossa on suomalaisten miesten ja naisten elinajanodotteet vuosilta 1966–2005 (lähde: Tilastokeskus).

Vuodet	Miehet	Naiset
1966–1970	65,9	73,6
1971–1975	66,7	75,2
1976–1980	68,5	77,2
1981–1985	70,1	78,4
1986–1990	70,7	78,8
1991–1995	72,1	79,7
1996–2000	73,5	80,8
2001–2005	75,1	81,9

a) Piirrä kuvaajat.
b) Laadi lineaaristen trendien yhtälöt.
c) Laadi vuosien 2006–2010 ennusteet.
d) Milloin naisten elinajanodote on 100 vuotta? Mieti vastausta!

14. Seuraavassa taulukossa esitetään patenttihakemusten ja myönnettyjen patenttien tilasto vuosilta 1991–2005 (lähde: Tilastokeskus).

Vuosi	Hakemuksia	Myönnetyt patentit
1991	6 191	2 681
1992	5 998	2 695
1993	5 975	2 721
1994	6 188	2 563
1995	6 364	2 347
1996	5 313	2 302
1997	4 668	2 298
1998	2 858	2 064
1999	2 870	1 786
2000	2 913	1 939
2001	2 668	2 052
2002	2 372	2 061
2003	2 187	2 402
2004	2 225	2 084
2005	2 061	1 757

Piirrä molempien aikasarjojen kuvaajat samaan kuvioon. Laadi vuoden 2015 ennusteet molemmille aikasarjoille. Mieti ennusteiden mielekkyyttä.

15. Laadi tehtävän 9 tilastoaineistosta sähköenergian hankintaa koskeva ennuste vuoden 2007 jokaisen kuukauden osalta.

16. Tutki edellisessä tehtävässä laaditun ennusteen hyvyyttä.

TEHTÄVIEN VASTAUKSIA

Luku 1

1. Ei

3. **a)** suhde, **b)** suhde, **c)** suhde, **d)** välimatka, **e)** luokittelu, **f)** luokittelu, **g)** järjestys, **h)** luokittelu, **i)** luokittelu, **j)** luokittelu, **k)** luokittelu, **l)** luokittelu, **m)** suhde, **n)** järjestys **o)** suhde.

4. **a)** Kyllä, **b)** Ei

Luku 2

1. Kyllä

7. **a)** 65, **b)** 25

8. A1: 120, A2: 300, A3: 480, A4: 360, A5: 240.

11. Ei.

12. Taulukon mukaan otoskoko on 1 500.

13. 1,9 prosenttiyksikköä.

14. Tuloksissa ei ole keskenään selvää ristiriitaa.

15. Ei, vaikka kumpikaan tutkimuslaitos ei osannut ennakoida oikein kolmen suurimman puolueen järjestystä.

16. Poliitikon väite saa tukea laskelmista.

Luku 3

–

Luku 4

1. Md = 13,8 %; $\overline{x}$ = 13,6 %; s = 1,26 % vaihteluväli = (11,3 %; 15,7 %)

2. 7,5

3. **a)** 10,5, **b)** 10,9

4. **a)** $\overline{x}$ = 36 000 km/v
b) s = 11 900 km/v
c) Md = 35 000 km/v
d) Eniten ajanut neljännes autoilijoista ajoi vähintään 41 500 km/v.

5. **a)** 5,6 pv, **b)** 5 pv, **c)** 4 pv, **d)** 2 pv, **e)** Q_1 = 2 pv, Q_3 = 8 pv, **f)** (0 pv, 19 pv), **g)** 1,15, **h)** 0,61

6. **a)** 5,7 pv, **b)** 4,4 pv

7.

8. Palkkojen keskiarvo on 29 700 €/v, mediaani 27 000 €/v ja moodi 29 000 €/v. Palkkojen keskihajonta on 13 500 €/v ja vaihteluväli (10 000 €/v; 73 000 €/v). Jakauma on oikealle vino, koska vinous = 1,22. Kvartiilit ovat: Q_1 = 20 000 €/v; Q_3 = 34 000 €/v.

10. **a)** 2,7 henkilöä,
b) 2 henkilöä,
c) 1 henkilö

11. **a)** ammattikorkeakoulututkinto
b) 250–499 €/kk

13. **a)** 4,1 mm, **b)** 4,1 mm,
c) 1,5 mm, **d)** 0,37 = 37 %

14. **a)** 3,1 mm, **b)** 2,0 mm,
c) (3,1 mm; 5,2 mm)

15. $\frac{R_1}{\overline{x}_1} = \frac{0{,}52\ V}{4{,}61\ V} \approx 0{,}113 = 11{,}3\ \%$

$\frac{R_2}{\overline{x}_2} = \frac{0{,}47\ V}{4{,}59\ V} \approx 0{,}102 = 10{,}2\ \%$

Luku 5

1. 720

2. **a)** 40 320,
b) 479 001 600,
c) $1{,}307674368 \cdot 10^{12}$

3. 150

4. 4 536

5. 180

6. 210

7. **a)** 32, **b)** 3003

8. 60 060

9. 63

10. Viking-lotossa. Siinä erilaisten ruudukoiden määrä on 12 271 512. Suomalaisessa lotossa erilaisia ruudukoita on 15 380 937.

11. 96 €

12. 2 598 960

13. 19 600

14. **a)** $\frac{1}{4}$, **b)** $\frac{1}{26}$

15. $\frac{3}{8}$

16. **a)** $\frac{16}{23}$, **b)** $\frac{13}{23}$

17. $\frac{7}{36}$

18. Todennäköisyydet ovat $\frac{5}{12}$ ja $\frac{1}{12}$.

19. **a)** $\frac{1}{2}$, **b)** $\frac{1}{4}$

20. **a)** 0,269, **b)** 0,163, **c)** 0,322

21. Keskimääräistä huonompaa, koska virheettömien suhteellinen osuus oli 0,994.

22. Noin 4,3 %

23. 0,594

24. **a)** 0,513, **b)** 0,487

25. **a)** 0,480, **b)** 0,632

26. **a)** 0,9025, **b)** 0,599

27. 0,094

28. **a)** 0,582, **b)** 0,238, **c)** 0,128, **d)** 0,052

29. **a)** 0,0000001, **b)** 0,000001

30. **a)** 0,00000063, **b)** 0,000016

31. **a)** 0,703, **b)** 0,274

32. **a)** 0,179, **b)** 0,0179, **c)** 0,179

34. **a)** 0,857, **b)** 0,045

35. 0,143

36. **a)** 0,444, **b)** 0,556

37. 0,665

38. **a)** 0,163, **b)** 0,99999995

39. **a)** 0,026, **b)** 0,049

40. **a)** 0,087, **b)** 0,999976

Luku 6

2. **a)** 0,75, **b)** 0,5

4. **a)** 0,784, **b)** 0,973, **c)** 0,657

5. **a)** $\frac{1}{6}$, **b)** $\frac{5}{6}$, **c)** $\frac{5}{18}$

6. 1,80 €

7. Noin 130 €

8. 224 €

10. **a)** 0,276, **b)** 0,820

11. 1,5

12. 600 kpl/v

13. 520 kpl/v

14. **a)** 0,294, **b)** 0,797, **c)** 0,497

15. **a)** 0,049, **b)** 0,042

16. **a)** 0,113, **b)** 0,029, **c)** 0,997

17. **a)** 0,375, **b)** 0,751, **c)** 0,625

18. **a)** 0,268, **b)** 0,891, **c)** 0,036

19. 0,185

20. 0,185

23. **a)** 0,7939, **b)** 0,2061, **c)** 0,950

24. **a)** –0,18, **b)** 0,18, **c)** 1,96

25. **a)** 0,603, **b)** 0,618, **c)** 0,612

26. $\mu = 69{,}6$

27. $\sigma = 3{,}85$

28. **a)** 0,348, **b)** 0,483

29. **a)** 0,789, **b)** 0,576

30. **a)** 0,184 , **b)** Noin 30 renkaan

31. **a)** 2,5 v

32. 0,097

33. 40 000 kilometrin välein

34. Nämä tulpat kestävät 40 000 kilometrin ajon todennäköisyydellä 0,9999998 ja edellisen tehtävän tulpat todennäköisyydellä 0,999.

35. 5 %:n todennäköisyydellä

36. 0,328

37. 4–5 päivänä vuodessa

38. 0,053

39. 0,017

40. Enintään 3 minuuttia

41. 5,5 %

42. 1 013 g

43. 0,40

Luku 7

1. **a)** 23,34–26,66;
b) 22,81–27,19;
c) 22,21–27,79

2. **a)** 90–102 tuntia vuodessa,
b) 85–107 tuntia vuodessa

3. 89,966 mm – 89,968 mm

4. 4,53 V – 4,59 V

5. 12 kuukauden takuu

6. Väliestimaatit ovat 22,8 kertaa – 24,2 kertaa ja 25,6 kertaa – 26,8 kertaa, joten pesunkestävyyden voidaan päätellä parantuneen.

7. 4,9 % – 7,7 %

8. 3 500 kpl – 4 920 kpl

9. 16,9 % – 23,5 %

10. **a)** Noin 2 100–2 400,
b) 48 €

11. 210 kpl – 300 kpl

12. **a)** 0,34 milj. € – 1,22 milj. €,
b) 9,40 €/kk eli noin 113 €/v

13. Luottamusvälit ovat:
A 32,0 % – 39,7 %; B 29,2 % – 36,8 %.
Tämän mukaan ei voida päätellä, että tuotteen A markkinaosuus olisi suurempi kuin tuotteen B markkinaosuus.

14. Noin 4 500

15. Z = –1,60. Levyjen laadun ei voida päätellä huonontuneen.

16. Z = –5,00. Tikkujen määrän voidaan päätellä vähentyneen.

17. Z = –2,18. Nollahypoteesia ei hylätä.

18. t = 2,44. Nollahypoteesi hylätään 5 %:n merkitsevyystasolla, mutta ei 1 %:n merkitsevyystasolla.

19. t = –1,25. Valmistajan väitettä ei voida kumota.

20. t = –0,65. Nollahypoteesia ei hylätä.

21. t = 2,82. Nollahypoteesi hylätään 5 %:n merkitsevyystasolla, mutta ei 1 %:n merkitsevyystasolla.

22. Z = –5,70. Tekstiilien pesunkestävyyden voidaan päätellä parantuneen.

23. Z = 3,19. Nollahypoteesi hylätään.

24. Z = 4,62. Nollahypoteesi hylätään.

25. Z = –0,91. Nollahypoteesia ei hylätä.

26. t = –1,96. Nollahypoteesi hylätään.

27. t = –0,31. Nollahypoteesia ei hylätä.

28. t = –4,76. Nollahypoteesi hylätään.

29. Z = –2,43. Nollahypoteesi hylätään sekä 5 %:n että 1 %:n merkitsevyystasolla.

30. Z = –4,67. Nollahypoteesi hylätään sekä 1 %:n että 0,1 %:n merkitsevyystasolla.

31. Z = –0,54. Nollahypoteesia ei hylätä.

32. Z = 1,78. Nollahypoteesia ei hylätä.

33. Z = –0,61. Nollahypoteesia ei hylätä.

34. Z = –0,72. Nollahypoteesia ei hylätä.

35. Z = 2,14. Nollahypoteesia ei hylätä.

36. Käytetään Mann–Whitneyn testiä.
U = –1,72. Nollahypoteesi hylätään yksisuuntaisessa testissä.

37. U = 0,30. Nollahypoteesia ei hylätä.

38. Käytetään Wilcoxonin testiä.
W = –3,13. Nollahypoteesi hylätään.

39. U = 1,13. Nollahypoteesia ei hylätä.

40. W = –3,27. Nollahypoteesi hylätään.

41. χ^2 = 7,80. Nollahypoteesia ei hylätä.

42. χ^2 = 5,50. Nollahypoteesia ei hylätä 1 %:n merkitsevyystasolla.

43. χ^2 = 1,27. Nollahypoteesia ei hylätä.

44. χ^2 = 6,16. Nollahypoteesia ei hylätä.

45. χ^2 = 2,16. Nollahypoteesia ei hylätä.

46. χ^2 = 22,57. Nollahypoteesi hylätään.

47. χ^2 = 55,94. Nollahypoteesi hylätään.

Luku 8

1. **a)** Ei., **b)** Kyllä.

2. Kyllä.

3. Ei.

6. Ei.

7. **a)** – **b)** 0,617

8. **a)** – **b)** 0,074

9. **a)** – **b)** 0,025

10. **a)** – **b)** 0,702

11. Hajontakuviot ja korrelaatiokertoimet vaihtelevat puhelinluettelosta löytyvien lukujen mukaan.

12. **a)** –

b) Hinnan ja pakastustehon välinen korrelaatiokerroin on 0,959. Hinnan ja bruttotilavuuden välinen korrelaatiokerroin on 0,973. Pakastustehon ja bruttotilavuuden välinen korrelaatiokerroin on 0,908.

13. Tulokset esitetään tehtävien mukaisesti:

Tehtävä 7: Bruttokansantulolla ja elinajanodotteella on todennäköisesti jokin yhteys keskenään.

Tehtävä 8: Avioliittojen solmimisella ja syntyneiden lasten lukumäärällä ei voida osoittaa olevan mitään yhteyttä toisiinsa.

Tehtävä 9: Imeväiskuolleisuudella ja henkilöiden lukumäärällä per lääkäri ei voida osoittaa olevan mitään yhteyttä toisiinsa.

Tehtävä 10: Hinnalla ja linkous-nopeudella on todennäköisesti jokin yhteys keskenään.

Tehtävä 11: Puhelinnumeroista muodostettujen muuttujien välillä ei pitäisi olla tilastollista yhteyttä.

14. Lehtien esittämät paremmuus-järjestykset ovat samansuuntaisia.

15. Henkilöiden mielipiteet ovat samansuuntaisia.

16. Työntekijöiden ja työnantajien mielipiteissä ei voida osoittaa olevan eroja käsiteltävien asioiden tärkeysjärjestyksestä.

17. Taustatietojen antaman kuvan ja haastattelun antaman kuvan välillä vallitsee yhteneväisyyttä.

18. **a)** Valaistuksella on merkitystä työsuoritukseen.

b) Valaistuksella on merkitystä työsuoritukseen.

c) Johtopäätös on **a)**- ja **b)**-kohdissa sama, mutta korrelaatiokertoimien arvoissa on hieman eroavaisuutta.

Luku 9

1. **a)** Tilastollinen
b) Funktionaalinen
c) Tilastollinen
d) Funktionaalinen
e) Tilastollinen

2. **a)** Kyllä, **b)** 1 000 €, **c)** 990 €.

3. Varauksellisesti:

CO_2-päästöt = 4,14 + 0,000198 · Bkt.

CO_2-päästöt = 9,1 tn/henkilö.

4. **a)** Kyllä
b) Osinko
= 0,175 + 0,0309 · Pörssikurssi.
Ennuste: 1,10 €.
c) 50 %
d) 59,9 %

5. **a)** –
b) Pituus
= 30017,4 + 0,333 · Lämpötila.
c) –
d) 135 m.

7. **a)** –
b) Koko = 5,44 + 0,233 · Hinta.
Ennuste: 191 000 €.
c) Ennuste: 87 m^2
d) 81,9 %
e) 20 m^2 – 150 m^2.

8. Väestö
= –1688,4 + 77,333 · (Vuosiluku-1900).
Selityskerroin on 99,9 %.

9. Elinajanodote
= 68,76 + 0,4416 · Bkt – 0,004280 · Bkt^2.
Selityskerroin on 79 %.

10. 77,1 vuotta. Suomen todellinen arvo on 78,5 vuotta.

11. **a)** Väestö
= –668,4 + 51,8333 · (Vuosiluku-1900)
+ 0,154545 · $(\text{Vuosiluku}-1900)^2$.
Ennuste vuodelle 2010: 6903 miljoonaa.
b) –
c) Ennusteet vuodelle 1950.

Mallina suora: 2 178 miljoonaa.

Mallina toisen asteen käyrä:
2 310 miljoonaa.

Mallina eksponenttifunktio:
2 626 miljoonaa.

12. **a)** Hinta = 33,12 – 1,664 · Ikä.
b) Hinta
= 34,21 – 0,100 · (Ajettu matk**a)**.
c) Hinta
= 34,44 –1,139 · Ikä – 0,0413 · (Ajettu matka).
d) 27 700 €.
e) –

14. **a)** Kulutus
= 2,917 + 0,0008766 · Tilavuus + 0,002128 · Omamassa.
b) Selityskerroin on 64 %.
c) –
d) –
e) Vertailu osoittaa, että ennusteet ovat keskimäärin tarkempia kuin Case C:n ennusteet.
f) Rajaus tuntui järkevältä.

Luku 10

4. Öljy: $y = 352\,400 + 1400{,}8 \cdot t$
(t = 1 vuonna 1995)

Hiili: $y = 165\,209 + 2364{,}2 \cdot t$
(t = 1 vuonna 1995)

Maakaasu:
$y = 116\,758 + 4217{,}4 \cdot t$
(t = 1 vuonna 1995)

Ydinenergia: $y = 207\,318 + 3447{,}6 \cdot t$
(t = 1 vuonna 1995)

5. Yhtälö: $86{,}789 + 7{,}9427 \cdot t$
(t = 1 vuonna 1996)

6. Aikasarjassa on todennäköisesti nouseva lineaarinen trendi.

7. Aikasarjassa on todennäköisesti nouseva lineaarinen trendi.

8. Ei voida osoittaa aikasarjassa olevan lineaarista nousevaa trendiä.

9. $y = 15390 + 144{,}65 \cdot t$
(t = 1 tammikuussa 2003)

Aikasarjassa on todennäköisesti nouseva lineaarinen trendi.

10. Lopulliset kausivaihteluindeksit:

1,291	1,180	1,193	0,934	0,845
0,714	0,689	0,774	0,827	0,976
1,232	1,345			

11. Trendisuoran yhtälö:
$y = 67{,}071 + 6{,}7237 \cdot t$
(t = 1 vuonna 1996)

Varauksellisesti ennuste: 168

12. Lineaarinen malli ei sovellu lainkaan ennusteiden laatimiseen.

Käyräviivainen kasvumalli:

$Y = e^{(-1{,}676 + 0{,}08232 \cdot t)}$

Ennuste 80-vuotiaalle: 136

13. **a)** –

b) Miehet:
$y = 64{,}443 + 1{,}3071 \cdot t$
($t = 1$ vuosina 1966–1970)

Naiset: $y = 73{,}164 + 1{,}119 \cdot t$
($t = 1$ vuosina 1966–1970)

c) Ennusteet vuosille 2006–2010:

Miehet: 76,2 vuotta

Naiset: 83,2 vuotta.

d) Ennustetta ei ole mielekästä laatia.

14. Ennusteita ei ole mielekästä laatia, sillä vuosi 2015 on liian kaukana havaintoarvoista.

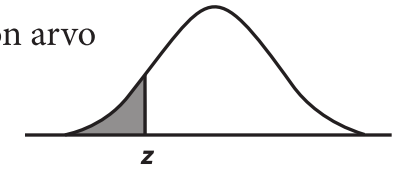
15. Lasketaan tulokset taulukkoon

Tammikuu	29 019
Helmikuu	26 695
Maaliskuu	27 161
Huhtikuu	21 400
Toukokuu	19 483
Kesäkuu	16 566
Heinäkuu	16 085
Elokuu	18 182
Syyskuu	19 546
Lokakuu	23 209
Marraskuu	29 475
Joulukuu	32 373

16. Malli on kohtuullisen hyvä. Tämän perusteella voidaan olettaa, että myös ennustearvot ovat varsin luotettavia.

NORMITETTU NORMAALIJAKAUMA

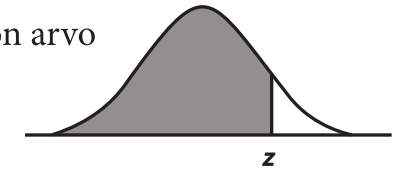
Normitetun normaalijakauman kertymäfunktion arvo pisteessä z vastaa varjostetun alueen pinta-alaa.



z	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
–3,4	,0003	,0003	,0003	,0003	,0003	,0003	,0003	,0003	,0003	,0002
–3,3	,0005	,0005	,0005	,0004	,0004	,0004	,0004	,0004	,0004	,0003
–3,2	,0007	,0007	,0006	,0006	,0006	,0006	,0006	,0005	,0005	,0005
–3,1	,0010	,0009	,0009	,0009	,0008	,0008	,0008	,0008	,0007	,0007
–3,0	,0013	,0013	,0013	,0012	,0012	,0011	,0011	,0011	,0010	,0010
–2,9	,0019	,0018	,0018	,0017	,0016	,0016	,0015	,0015	,0014	,0014
–2,8	,0026	,0025	,0024	,0023	,0023	,0022	,0021	,0021	,0020	,0019
–2,7	,0035	,0034	,0033	,0032	,0031	,0030	,0029	,0028	,0027	,0026
–2,6	,0047	,0045	,0044	,0043	,0041	,0040	,0039	,0038	,0037	,0036
–2,5	,0062	,0060	,0059	,0057	,0055	,0054	,0052	,0051	,0049	,0048
–2,4	,0082	,0080	,0078	,0075	,0073	,0071	,0069	,0068	,0066	,0064
–2,3	,0107	,0104	,0102	,0099	,0096	,0094	,0091	,0089	,0087	,0084
–2,2	,0139	,0136	,0132	,0129	,0125	,0122	,0119	,0116	,0113	,0110
–2,1	,0179	,0174	,0170	,0166	,0162	,0158	,0154	,0150	,0146	,0143
–2,0	,0228	,0222	,0217	,0212	,0207	,0202	,0197	,0192	,0188	,0183
–1,9	,0287	,0281	,0274	,0268	,0262	,0256	,0250	,0244	,0239	,0233
–1,8	,0359	,0351	,0344	,0336	,0329	,0322	,0314	,0307	,0301	,0294
–1,7	,0446	,0436	,0427	,0418	,0409	,0401	,0392	,0384	,0375	,0367
–1,6	,0548	,0537	,0526	,0516	,0505	,0495	,0485	,0475	,0465	,0455
–1,5	,0668	,0655	,0643	,0630	,0618	,0606	,0594	,0582	,0571	,0559
–1,4	,0808	,0793	,0778	,0764	,0749	,0735	,0721	,0708	,0694	,0681
–1,3	,0968	,0951	,0934	,0918	,0901	,0885	,0869	,0853	,0838	,0823
–1,2	,1151	,1131	,1112	,1093	,1075	,1056	,1038	,1020	,1003	,0985
–1,1	,1357	,1335	,1314	,1292	,1271	,1251	,1230	,1210	,1190	,1170
–1,0	,1587	,1562	,1539	,1515	,1492	,1469	,1446	,1423	,1401	,1379
–0,9	,1841	,1814	,1788	,1762	,1736	,1711	,1685	,1660	,1635	,1611
–0,8	,2119	,2090	,2061	,2033	,2005	,1977	,1949	,1922	,1894	,1867
–0,7	,2420	,2389	,2358	,2327	,2296	,2266	,2236	,2206	,2177	,2148
–0,6	,2743	,2709	,2676	,2643	,2611	,2578	,2546	,2514	,2483	,2451
–0,5	,3085	,3050	,3015	,2981	,2946	,2912	,2877	,2843	,2810	,2776
–0,4	,3446	,3409	,3372	,3336	,3300	,3264	,3228	,3192	,3156	,3121
–0,3	,3821	,3783	,3745	,3707	,3669	,3632	,3594	,3557	,3520	,3483
–0,2	,4207	,4168	,4129	,4090	,4052	,4013	,3974	,3936	,3897	,3859
–0,1	,4602	,4562	,4522	,4483	,4443	,4404	,4364	,4325	,4286	,4247
–0,0	,5000	,4960	,4920	,4880	,4840	,4801	,4761	,4721	,4681	,4641

NORMITETTU NORMAALIJAKAUMA

Normitetun normaalijakauman kertymäfunktion arvo pisteessä z vastaa varjostetun alueen pinta-alaa.



z	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0,0	,5000	,5040	,5080	,5120	,5160	,5199	,5239	,5279	,5319	,5359
0,1	,5398	,5438	,5478	,5517	,5557	,5596	,5636	,5675	,5714	,5753
0,2	,5793	,5832	,5871	,5910	,5948	,5987	,6026	,6064	,6103	,6141
0,3	,6179	,6217	,6255	,6293	,6331	,6368	,6406	,6443	,6480	,6517
0,4	,6554	,6591	,6628	,6664	,6700	,6736	,6772	,6808	,6844	,6879
0,5	,6915	,6950	,6985	,7019	,7054	,7088	,7123	,7157	,7190	,7224
0,6	,7257	,7291	,7324	,7357	,7389	,7422	,7454	,7486	,7517	,7549
0,7	,7580	,7611	,7642	,7673	,7704	,7734	,7764	,7794	,7823	,7852
0,8	,7881	,7910	,7939	,7967	,7995	,8023	,8051	,8078	,8106	,8133
0,9	,8159	,8186	,8212	,8238	,8264	,8289	,8315	,8340	,8365	,8389
1,0	,8413	,8438	,8461	,8485	,8508	,8531	,8554	,8577	,8599	,8621
1,1	,8643	,8665	,8686	,8708	,8729	,8749	,8770	,8790	,8810	,8830
1,2	,8849	,8869	,8888	,8907	,8925	,8944	,8962	,8980	,8997	,9015
1,3	,9032	,9049	,9066	,9082	,9099	,9115	,9131	,9147	,9162	,9177
1,4	,9192	,9207	,9222	,9236	,9251	,9265	,9279	,9292	,9306	,9319
1,5	,9332	,9345	,9357	,9370	,9382	,9394	,9406	,9418	,9429	,9441
1,6	,9452	,9463	,9474	,9484	,9495	,9505	,9515	,9525	,9535	,9545
1,7	,9554	,9564	,9573	,9582	,9591	,9599	,9608	,9616	,9625	,9633
1,8	,9641	,9649	,9656	,9664	,9671	,9678	,9686	,9693	,9699	,9706
1,9	,9713	,9719	,9726	,9732	,9738	,9744	,9750	,9756	,9761	,9767
2,0	,9772	,9778	,9783	,9788	,9793	,9798	,9803	,9808	,9812	,9817
2,1	,9821	,9826	,9830	,9834	,9838	,9842	,9846	,9850	,9854	,9857
2,2	,9861	,9864	,9868	,9871	,9875	,9878	,9881	,9884	,9887	,9890
2,3	,9893	,9896	,9898	,9901	,9904	,9906	,9909	,9911	,9913	,9916
2,4	,9918	,9920	,9922	,9925	,9927	,9929	,9931	,9932	,9934	,9936
2,5	,9938	,9940	,9941	,9943	,9945	,9946	,9948	,9949	,9951	,9952
2,6	,9953	,9955	,9956	,9957	,9959	,9960	,9961	,9962	,9963	,9964
2,7	,9965	,9966	,9967	,9968	,9969	,9970	,9971	,9972	,9973	,9974
2,8	,9974	,9975	,9976	,9977	,9977	,9978	,9979	,9979	,9980	,9981
2,9	,9981	,9982	,9982	,9983	,9984	,9984	,9985	,9985	,9986	,9986
3,0	,9987	,9987	,9987	,9988	,9988	,9989	,9989	,9989	,9990	,9990
3,1	,9990	,9991	,9991	,9991	,9992	,9992	,9992	,9992	,9993	,9993
3,2	,9993	,9993	,9994	,9994	,9994	,9994	,9994	,9995	,9995	,9995
3,3	,9995	,9995	,9995	,9996	,9996	,9996	,9996	,9996	,9996	,9997
3,4	,9997	,9997	,9997	,9997	,9997	,9997	,9997	,9997	,9997	,9998

T-JAKAUMAAN LIITTYVIÄ KRIITTISIÄ ARVOJA

f	Merkitsevyystaso yksisuuntaisessa testissä							
	10 %	5 %	2,5 %	1 %	0,5 %	0,1 %	0,05 %	0,01 %
	Merkitsevyystaso kaksisuuntaisessa testissä							
	20 %	10 %	5 %	2 %	1 %	0,2 %	0,1 %	0,02 %
1	3,078	6,314	12,706	31,821	63,656	318,29	636,58	3185,3
2	1,886	2,920	4,303	6,965	9,925	22,328	31,600	70,706
3	1,638	2,353	3,182	4,541	5,841	10,214	12,924	22,203
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610	13,039
5	1,476	2,015	2,571	3,365	4,032	5,894	6,869	9,676
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959	8,023
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408	7,064
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041	6,442
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781	6,009
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587	5,694
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437	5,453
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318	5,263
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221	5,111
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140	4,985
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073	4,880
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015	4,790
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965	4,715
18	1,330	1,734	2,101	2,552	2,878	3,610	3,922	4,648
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883	4,590
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850	4,539
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819	4,492
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792	4,452
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768	4,416
24	1,318	1,711	2,064	2,492	2,797	3,467	3,745	4,382
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725	4,352
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707	4,324
27	1,314	1,703	2,052	2,473	2,771	3,421	3,689	4,299
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674	4,276
29	1,311	1,699	2,045	2,462	2,756	3,396	3,660	4,254
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646	4,234
40	1,303	1,684	2,021	2,423	2,704	3,307	3,551	4,094
50	1,299	1,676	2,009	2,403	2,678	3,261	3,496	4,014
60	1,296	1,671	2,000	2,390	2,660	3,232	3,460	3,962
80	1,292	1,664	1,990	2,374	2,639	3,195	3,416	3,899
100	1,290	1,660	1,984	2,364	2,626	3,174	3,390	3,861
150	1,287	1,655	1,976	2,351	2,609	3,145	3,357	3,813
200	1,286	1,653	1,972	2,345	2,601	3,131	3,340	3,789
300	1,284	1,650	1,968	2,339	2,592	3,118	3,323	3,765
500	1,283	1,648	1,965	2,334	2,586	3,107	3,310	3,747
∞	1,282	1,645	1,960	2,326	2,576	3,090	3,290	3,719

χ^2-JAKAUMAN KRIITTISIÄ ARVOJA

	Merkitsevyystaso yksisuuntaisessa testissä							
f	**99,9 %**	**99,5 %**	**99 %**	**95 %**	**5 %**	**1 %**	**0,5 %**	**0,1 %**
1	0,000	0,000	0,000	0,004	3,841	6,635	7,879	10,827
2	0,002	0,010	0,020	0,103	5,991	9,210	10,597	13,815
3	0,024	0,072	0,115	0,352	7,815	11,345	12,838	16,266
4	0,091	0,207	0,297	0,711	9,488	13,277	14,860	18,467
5	0,210	0,412	0,554	1,145	11,070	15,086	16,750	20,515
6	0,381	0,676	0,872	1,635	12,592	16,812	18,548	22,457
7	0,599	0,989	1,239	2,167	14,067	18,475	20,278	24,321
8	0,857	1,344	1,647	2,733	15,507	20,090	21,955	26,124
9	1,152	1,735	2,088	3,325	16,919	21,666	23,589	27,877
10	1,479	2,156	2,558	3,940	18,307	23,209	25,188	29,588
11	1,834	2,603	3,053	4,575	19,675	24,725	26,757	31,264
12	2,214	3,074	3,571	5,226	21,026	26,217	28,300	32,909
13	2,617	3,565	4,107	5,892	22,362	27,688	29,819	34,527
14	3,041	4,075	4,660	6,571	23,685	29,141	31,319	36,124
15	3,483	4,601	5,229	7,261	24,996	30,578	32,801	37,698
16	3,942	5,142	5,812	7,962	26,296	32,000	34,267	39,252
17	4,416	5,697	6,408	8,672	27,587	33,409	35,718	40,791
18	4,905	6,265	7,015	9,390	28,869	34,805	37,156	42,312
19	5,407	6,844	7,633	10,117	30,144	36,191	38,582	43,819
20	5,921	7,434	8,260	10,851	31,410	37,566	39,997	45,314
21	6,447	8,034	8,897	11,591	32,671	38,932	41,401	46,796
22	6,983	8,643	9,542	12,338	33,924	40,289	42,796	48,268
23	7,529	9,260	10,196	13,091	35,172	41,638	44,181	49,728
24	8,085	9,886	10,856	13,848	36,415	42,980	45,558	51,179
25	8,649	10,520	11,524	14,611	37,652	44,314	46,928	52,619
26	9,222	11,160	12,198	15,379	38,885	45,642	48,290	54,051
27	9,803	11,808	12,878	16,151	40,113	46,963	49,645	55,475
28	10,391	12,461	13,565	16,928	41,337	48,278	50,994	56,892
29	10,986	13,121	14,256	17,708	42,557	49,588	52,335	58,301
30	11,588	13,787	14,953	18,493	43,773	50,892	53,672	59,703
40	17,917	20,707	22,164	26,509	55,758	63,691	66,766	73,403
50	24,674	27,991	29,707	34,764	67,505	76,154	79,490	86,660
60	31,738	35,534	37,485	43,188	79,082	88,379	91,952	99,608
80	46,520	51,172	53,540	60,391	101,879	112,329	116,321	124,839
100	61,918	67,328	70,065	77,929	124,342	135,807	140,170	149,449

PEARSONIN KORRELAATIOKERTOIMEN KRIITTISIÄ ARVOJA

Otoskoko	Merkitsevyystaso					
n	5 % 1 %	2,5 % 5 %	1 % 2 %	0,5 % 1 %	0,05 % 0,1 %	yksisuuntainen kaksisuuntainen
5	0,805	0,878	0,934	0,959	0,991	
6	0,729	0,811	0,882	0,917	0,974	
7	0,669	0,754	0,833	0,875	0,951	
8	0,621	0,707	0,789	0,834	0,925	
9	0,582	0,666	0,750	0,798	0,898	
10	0,549	0,632	0,715	0,765	0,872	
11	0,521	0,602	0,685	0,735	0,847	
12	0,497	0,576	0,658	0,708	0,823	
13	0,476	0,553	0,634	0,684	0,801	
14	0,458	0,532	0,612	0,661	0,780	
15	0,441	0,514	0,592	0,641	0,760	
16	0,426	0,497	0,574	0.623	0,742	
17	0,412	0,482	0,558	0,606	0.725	
18	0,400	0,468	0,543	0,590	0,708	
19	0,389	0,456	0,529	0,575	0,693	
20	0,378	0,444	0,516	0,561	0,679	
21	0,369	0,433	0,503	0,549	0,665	
22	0,360	0,423	0,492	0,537	0,652	
23	0,352	0,413	0,482	0,526	0,640	
24	0,344	0,404	0,472	0,515	0,629	
25	0,337	0,396	0,462	0,505	0,618	
26	0,330	0,388	0,453	0,496	0,607	
27	0,323	0,381	0,445	0,487	0,597	
28	0,317	0,374	0,437	0,479	0,588	
29	0,311	0,367	0,430	0,471	0,579	
30	0,306	0,361	0,423	0,463	0,570	
40	0,264	0,312	0,367	0,403	0,501	
50	0,235	0,279	0,328	0,361	0,451	
60	0,214	0,254	0,300	0,330	0,414	
70	0,198	0,235	0,278	0,306	0,385	
80	0,185	0,220	0,260	0,286	0,361	
90	0,174	0,207	0,245	0,270	0,341	
100	0,165	0,197	0,232	0,256	0,324	
200	0,117	0,139	0,164	0,182	0,231	
300	0,095	0,113	0,134	0,149	0,189	
400	0,082	0,098	0,116	0,129	0,164	
500	0,074	0,088	0,104	0,115	0,147	
750	0,060	0,072	0,085	0,094	0,120	
1000	0,052	0,062	0,074	0,081	0,104	

SPEARMANIN JÄRJESTYSKORRELAATIOKERTOIMEN KRIITTISIÄ ARVOJA

Otoskoko	Merkitsevyystaso				
n	5 % 10 %	2,5 % 5 %	1 % 2 %	0,5 % 1 %	yksisuuntainen kaksisuuntainen
5	0,900	–	–	–	
6	0,829	0,886	0,943	–	
7	0,714	0,786	0,893	0,929	
8	0,643	0,738	0,833	0,881	
9	0,600	0,700	0,783	0,833	
10	0,564	0,648	0,745	0,794	
11	0,536	0,618	0,709	0,755	
12	0,503	0,587	0,678	0,727	
13	0,484	0,560	0,648	0,703	
14	0,464	0,538	0,626	0,679	
15	0,446	0,521	0,604	0,654	
16	0,429	0,503	0,582	0,635	
17	0,414	0,488	0,566	0,618	
18	0,401	0,472	0,550	0,600	
19	0,391	0,460	0,535	0,584	
20	0,380	0,447	0,522	0,570	
21	0,370	0,436	0,509	0,556	
22	0,361	0,425	0,497	0,544	
23	0,353	0,416	0,486	0,532	
24	0,344	0,407	0,477	0,521	
25	0,337	0,398	0,466	0,511	
26	0,331	0,390	0,457	0,501	
27	0,324	0,383	0,449	0,491	
28	0,318	0,375	0,440	0,483	
29	0,312	0,368	0,433	0,477	
30	0,306	0,362	0,425	0,467	
35	0,283	0,335	0,394	0,433	
40	0,264	0,313	0,368	0,405	
45	0,248	0,294	0,347	0,382	
50	0,235	0,279	0,329	0,363	

HAKEMISTO

A

adjusted coefficient of determination 278
aikasarja 305
aikasarja-analyysi 305
aikasarjan komponentti 311
alakvarttiili 86
alkeistapahtuma 108
arithmetic mean 83
aritmeettinen keskiarvo 83
askeltava regressioanalyysi 282

B

backward elimination 279
bias 33
binomial distribution 132
binomijakauma 132
binomijakauman kertymäfunktio 135
binomijakauman parametri 133
binomijakauman
todennäköisyysfunktio 133

C

case-tutkimus 21
census 29
centered moving average 321
central limit theorem 150
cluster sampling 35
coefficient of correlation 233
coefficient of determination 277
coefficient of variation 92
confidence level 166
conjecture 14
convenience sampling 36
coverage 41
cross-sectional study 21
cumulative distribution function 135
cumulative probability 127
curve fitting 260
cyclical behavior 313

D

Danielsin testi 318
data mining 23
decile 86
deduktiivinen 14
deflatoida 307
degrees of freedom 182
density function 141
descriptive statistic 46
desiili 86
deterministinen malli 25
diagnostiikka 283
diagrammi 53
diskreetti 16
diskreetti todennäköisyysjakauma 126

E

edustava 29
eksponentiaalinen malli 269
eksponentiaalinen tasoitus 320
empiirinen 20
ennusteväli 289, 290
ennustevyöhyke 289
epäjatkuva 125
epäsäännöllinen vaihtelu 314
erillisten tapahtumien
yhteenlaskusääntö 114
esitutkimus 18
estimaatti 166
estimaattori 166
estimate 266
estimointi 165
etenevä valinta 282
Eurostat 23
experimental 21
exponential smoothing 320

F

forward selection 282
fractile 86
fraktiili 86
frame 30
frekvenssi 48
frekvenssitaulukko 48

G

Gaussin käyrä 144

H

hajonta 97
hajontakuvio 65, 229
hajontaluku 78, 88

harha 33, 336
harkinnanvarainen otanta 36
havainnoiva tutkimus 21
havaintomatriisi 46
havaittu frekvenssi 200
histogrammi 60
huipukkuus 94
hylkäämisalue 178
hylkäämisvirhe 176
hypoteesi 175
hypoteesien testaus 165, 175
hyväksymisvirhe 176

I

indeksi 307
induktiivinen 14
inference 14
inferential statistics 165
internet 23
Internet-tunnus 25
interpretation 19
interquartile range 97
interval scale 15
irregular movement 314

J

X^2-jakauma 156
jakaumasta riippumaton testi 195
jatkuva 16
jatkuvan jakauman kertymäfunktio 142
jatkuva todennäköisyysjakauma 141
johtopäätös 13
järjestysasteikko 15, 195, 125
järkevyystarkastelu 228
jäännöstermi 282

K

kahden riippumattoman otoksen keskiarvojen vertailu t-testin avulla 187
kahden tapahtuman yleinen yhteenlaskusääntö 115
kaksiulotteinen frekvenssijakauma 52
kato 41
kausaalinen tutkimus 21
kausipuhdistettu 321
kausivaihtelu 312, 319
kausivaihteluindeksi 322
kehikko 30
kertymäfunktio 135
kertymätodennäköisyys 127
keskeinen raja-arvolause 150
keskiarvo 97
keskiarvon keskivirhe 154
keskiarvotesti 178
keskihajonta 90, 97
keskiluku 78
keskitetty liukuva keskiarvo 321
keskivirhe 266
khi distribution 156
kiintiöotanta 36
kirjoituspöytätutkimus 21
klassinen todennäköisyys 108
kokeellinen tutkimus 21
kokonaistutkimus 29
Kolmogorov–Smirnovin testi 205
kombinaatio 106
komplementtitapahtuma 118
komplementtitapahtuman todennäköisyys 118
kontingenssitaulukko 202
korrelaatiodiagrammi 229
korrelaatiokerroin 233
korrelaatiokertoimen merkitsevyys 242
korrelaatiokertoimen ominaisuuksia 234
korrelaatiomatriisi 239
kriittinen arvo 168
kumulatiivinen frekvenssi 50
kurtosis 94
kuvaileva tilastotiede 46
kuvaileva tutkimus 21
kvalitatiivinen 16
kvantitatiivinen 16
kvartiili 86
kvartiilipoikkeama 90
kvartiiliväli 89
kyselylomake 42
käyrän sovitus 260

L

lineaarinen malli 261
liukuva keskiarvo 320
longitudinal study 21
luokitellun tilastoaineiston mediaani 81
luokitteluasteikko 15, 195
luottamusraja 166
luottamustaso 166

luottamusväli 166, 290
luvun kertoma 105
lähdekritiikki 20

M

malli 14
malliajattelu 259
mallin hyvyys 277
Mann–Whitneyn testi 197, 205
margin of error 38
matemaattinen malli 25
maximum 97
Md 80
mean 83
measures of central tendency 79
measures of location 78
mediaani 80
mediaanipalkka 97
median 80
merkitsevyystaso 177
minimum 97
Mo 79
mode 79
moodi 79
moving average 321
multicollinearity 275
multikollineaarisuus 275
muuttuja 16

N

nollahypoteesi 175
nominal scale 15
nonresponse 41
normaalijakauma 144
normaalijakauman kertymäfunktio 146
normal distribution 144
normitettu satunnaismuuttuja 146
näyte 29

O

observational 21
odotusarvo 130, 131
odotusarvon luottamusväli 168
OLS 259
optimum allocation 34
ordinal scale 15
ordinary least squares 259
osite 33
ositettu otanta 33
otaksuma 14
otantajakauma 154
otantatutkimus 29
otantatutkimuksen vaihe 30
otantavirhe 29
otantayksikkö 29
otos 29
otoskoko 164, 173
outlier 287

P

palvelutasoprofiili 209
parametrien estimointi 165
parametrinen testi 178, 205
paras kiintiö 34
parvikuvio 65
p-arvo 177, 242, 266, 278, 292
Pearsonin korrelaatiokerroin 233
Pearson's correlation coefficient 233
peitto 41
permutaatio 104
perusjoukko 15
perustutkimus 20
pienimmän neliösumman menetelmä 259
piste-estimaatti 166
pistetodennäköisyys 129
pitkittäistutkimus 21
poikkeava arvo 287
poikkileikkaustutkimus 21
poimintaväli 32
Poisson distribution 137
Poisson-jakauma 137
poistava valinta 279
polynomimalli 273
population 15
prediction band 290
prediction interval 290
probability function 133
prosenttiosuuden luottamusväli 170
prospektiivinen tutkimus 21
pylväsdiagrammi 53

Q

quartile 86, 97
quota sampling 36

R

random variable 125
range 89
raportti 20
ratio scale 16
regressand 261
regressioanalyysi 261
regressiokerroin 262, 265, 275
regression coefficients 265
regressiosuoran yhtälö 261
regressor 261
reliabiliteetti 17
research 18
residual 282
retrospektiivinen tutkimus 21
riippumattoman tapahtuman kertolaskusääntö 116
X^2-riippumattomuustesti 202
ristiintaulukointi 52
rypäs 35
ryväsotanta 35

S

samanaikainen hajontakuvio 230
sample 29
sample standard deviation 97
sampling error 29
sampling unit 29
satunnaisilmiö 107
satunnaiskoe 107
satunnaismuuttuja 125
satunnaisotanta 31
satunnaisvaihtelu 314, 327
scatter diagram 229
scatterplot 229
science 13
scientific method 13
seasonal fluctuation 312
sektoridiagrammi 63
selitettävä 261
selittävä muuttuja 261
selitysaste 277
selityskerroin 277
sijaintiluku 78
simple moving average 320
simple random sampling 31
skenaario 266
skewness 94
smoothing 320
soveltava tutkimus 20
Spearmanin järjestyskorrelaatiokerroin 239, 240
Spearman's coefficient of rank correlation 239, 240
standard deviation 90
standard error 266
StatFin 23
statistic 25
statistical probability 111
stepwise regression analysis 282
stokastinen malli 25
strata 33
stratified sampling 33
Studentin t-jakauma 155
Studentin t-testi 182
Student's distribution 155
study 18
suhdannevaihtelu 313, 325
suhdeasteikko 16
suhteellinen kiintiöinti 34
suhteellisen osuuden testaus 192
suhteellisen osuuden testaus kahdesta otoksesta 193
suhteutettu selityskerroin 278
summamalli 314
suora jakauma 47
survey-tutkimus 21
systemaattinen otanta 32
systemaattinen virhe 33
systemaattinen virhetilanne 16
systematic sampling 32

T

tapahtuma 108
t-arvo 266, 278, 292
tasainen kiintiö 34
tasoitus 320
taustamuuttuja 43
teemakartta 64
teoreettinen frekvenssi 200
X^2-testi 205, 214
testimuuttuja 179
testimuuttujan arvo 178
testin valinta 204

testin voimakkuus 205
tiede 13
tiedon louhinta 23
tieteellinen menetelmä 13
tietokanta 23
tietopankki 23
tiheysfunktio 141
Tilastokeskus 22
tilastollinen ajatteluprosessi 26
tilastollinen päättely 165
tilastollinen testi 175
tilastollinen todennäköisyys 111
tilastollinen vuosikirja 22
tilastollisesti erittäin merkitsevä 177
tilastollisesti melkein merkitsevä 177
tilastollisesti merkitsevä 177
tilastosymboli 23
tilastotiede 25
tilastoyksikkö 15
todennäköisyysfunktio 133
todennäköisyysjakauma 126
todennäköisyysmalli 125
tracking 320
trend 311
trendi 311, 316
t-testi 205
tulkinta 19
tulkintavirhe 247
tulomalli 314
tuloperiaate 104
tunnusluku 78
tutkimus 18
tutkimusongelma 18
tutkimusprosessi 18
tutkimusstrategia 18
tutkimusyksikkö 15
tyyppiarvo 79

V

vaihteluväli 89
vaihteluvälin pituus 90
vaihtoehtoinen hypoteesi 175
validiteetti 16
vapausaste 155, 132, 243
variaatiokerroin 92
variance 92
variance inflatory factor 279
varianssi 92
verrannollisten parien t-testi 190
VIF-arvo 279
viivadiagrammi 61
vinous 94, 97
virhemarginaali 38
visio 266
visuaalisen vertailtavuuden periaate 53
vuokaavio 248, 292
väliestimaatti 166
välimatka-asteikko 15, 178, 205

W

WebStat 23
Wilcoxonin testi 198, 205

X

X^2-jakauma 156
X^2-riippumattomuustesti 202
X^2-testi 205, 214
X^2-yhteensopivuustesti 200

Y

X^2-yhteensopivuustesti 200
yhteiskorrelaatiokerroin 278
yleinen kertolaskusääntö 117
yläkvartiili 86